CROWD-SOURCED USER EXPERIENCE EVALUATION AND ADAPTATION FOR A PERSONALIZED USER INTERFACE

Buddhima Naween Rathnayake

(179348J)

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

May 2019

CROWD-SOURCED USER EXPERIENCE EVALUATION AND ADAPTATION FOR A PERSONALIZED USER INTERFACE

Buddhima Naween Rathnayake

(179348J)

Thesis submitted in partial fulfillment of the requirements for the Degree Master of Science in Computer Science specializing in Software Architecture

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

May 2019

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgment to any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in partial print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

-------------------------------------   ---------------------------------

Signature   Date

The above candidate has carried out research for the Masters thesis under my supervision.

Name of the Supervisor: Dr. Indika Perera

-------------------------------------   ------------------------------------

Signature of the Supervisor   Date

ACKNOWLEDGEMENT

This study would not have been possible without the guidance and the help of several individuals who extended their valuable assistance in the preparation and completion of this dissertation. First and foremost,

I'm grateful to my supervisor, Dr. Indika Perera, for his guidance, patience and providing me with an excellent atmosphere for the completion of this research work.

I thank all the teaching and administrative staff of University of Moratuwa, for their service and support.

Great deals appreciated go to Dr. Charith Chitraranjan & Dr. Malaka Walpola, Course main lecturers for their guidance and contribution throughout the course.

Special thanks to my colleagues whose commitment and support contributed to this project and who helped me in numerous ways.

Finally, I would like to extend my regards to my family members who showered me with the needed power of confidence and encouragement provided to make this study a success.

ABSTRACT

We have been using a vast array of software products which provide different services. Even there are many products which fulfill the same requirement some software products stand out from the rest. Providing a strong user experience has been the driving factor for the success of a software product. The goal of the user experience design is to improve user satisfaction, loyalty, and ease of use. To maintain constant user satisfaction throughout the product lifetime, conducting user experience evaluations are desirable however these evaluations tend to be short-term evaluations only focusing on initial product designs. This research aims to promote crowdsourcing mechanisms to gather user feedback on the user experience of any website. These evaluations would be carried out through conversations based questionnaires and capture users perspective of the product. The system can be integrated into any website and it will gather users' answers and their sentiment through conversation-based questionnaires. These user evaluations are analyzed to determine what the user desire and user configuration will be saved as personalized content. These personalized content in the user interface are again to be used to represent the website based on the majority perspective. This research aims to provide insight on how conversational chat-bots are capable of capturing personas of the feedback providers and increasing the feedback rate than questionnaires. Based on the feedback gathered through the Chatbot users 73.1% of the users rated that they choose Chatbot over filling online forms. On how Chatbot is capable of mimicking a human being 12% of the users thought they were actually talking to a real human being while 68% thought it was a computer-based program. The application consists of an admin dashboard which represents demographic data, overall sentiment and sentiment score variation over time, which would be an immense help to the usability evaluation of a particular website.

Keywords: usability evaluation, crowdsourcing, conversational interfaces, sentiment analysis

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation		Description
UX	-	User Experience
UI	-	User Interface
EM	-	Expectation Maximization
NLP	-	Natural Language Processing
NLU	-	Natural Language Understanding
UUID	-	Universal Unique Identifier
API	-	Application Programming Interface

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

We have been using websites to access our favorite services, for e-commerce, as information portals, and for various data representations. However, the requirement is not just to represent data as processed information. The real benefit lies whether the users are convenient in using the system or the website. We call those systems are user-friendly. They promote performing user's tasks more conveniently than others. So as an example if we consider the e-commerce domain there are a variety of e-commerce websites which advertise various categories of products, assuming you are the consumer what would you choose to purchase the product you wanted, the user-friendly website or the other? It's quite obvious that people tend to choose convenience over other constraints. We try to capture this non-functional quality as usability which describes how easy a product can be used, but today it has grown up to overall user experience.

1.1 User Experience and Evaluation

User Experience (UX) refers to a person's overall experience of the system. It can be emotions people have or their attitude towards the system. Evaluation of user experience drives us to have a deep understanding of the users of the system, what they desire, what they need and what needed to be improved. The international standard on ergonomics of human system interaction, ISO 9241-210 defines user experience as "a person's perceptions and responses that result from the use or anticipated use of a product, system or service". So User Experience is a crucial factor to consider on product success, this is what differs some businesses perform better than the others even the services they provide are identical. According to Peter Morville's user experience honeycomb [2], there are seven factors which influence user experience. Those factors are,

- Useful
- Usable
- Findable
- Credible
- Desirable

- Accessible
- Valuable

Peter Morville explains that in order to have a better design and meaningful and valuable user experience information and the system itself should pursue those characters. Every facet of user experience is captured through these factors. Products that fulfill these factors tend to be successful in the market.
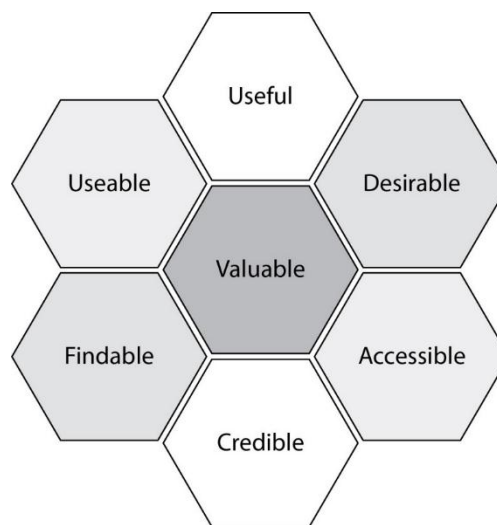
Figure 1: User Experience Honeycomb

Being useful is that the developed product or service should be useful to a user or the consumer and the service that the user expects is fulfilled accordingly. If the reason why the users use the system is not fulfilled there will be no purpose of using it. The system should be easy to use. This is captured through the usable property. Usability professionals sometimes come across people who are frustrated with software products, however, it is identified that they are not frustrated by the service the product provides, simply by the usability issues of the product [9]. The learning curve user has to go through to perform his/her tasks should be painless. In our context when a system is released people tend to get used to products they will discard ones with complex scenarios to perform the required tasks. However, there are many other dependencies which will affect this such as the service the system provides, how valuable is it to the user and in e-commerce, there will also be the factor of monetary value a product

represents. It's hard to define what actually made the system didn't turn around as expected, but we can come to the conclusion that without a pleasing design it surely won't. Being Desirable is capturing the user interface facet, the visual aesthetics of the product and having a minimalist design. A good example is google.com where home page only contains a textbox and two buttons, each performing crucial tasks of a search engine, nothing more. Findable is that the information is easy to navigate through. These navigational structures should be capable to get what the user need within a short period of time. Because this organization is how you are going to guide visitors through your material. Accessible describes that the product should be capable to provide services to the whole community including disabled people as well. (E.g. usage of "ALT" attributes on image sources). Credibility is that the users' trust in the system. Finally, the product should be valuable it must provide value to both the business and the consumers as well. It's examined that with the maturity of the industrial sector users tend to use products that provide engaging user experience over others, so having more focus on these aspects are desirable [7].

1.2 User Experience Elements

Garrett in his book [1] "The Elements of User Experience" defines user experience based on a simple scenario where the user tries to buy something online. Here we study the elements of the experience based on the actions user make. The five elements of UX is a combination of five dependent layers where at each level another layer will be built upon it, from bottom to top layers are organized to start with the abstract level to concrete level. This dependency between the layers is described by Garrett as "The choices you make on each plane affect the choices available to you on the next plane above it". So the decisions we take at the abstract level has a ripple effect of dependencies and will affect at the concrete level.
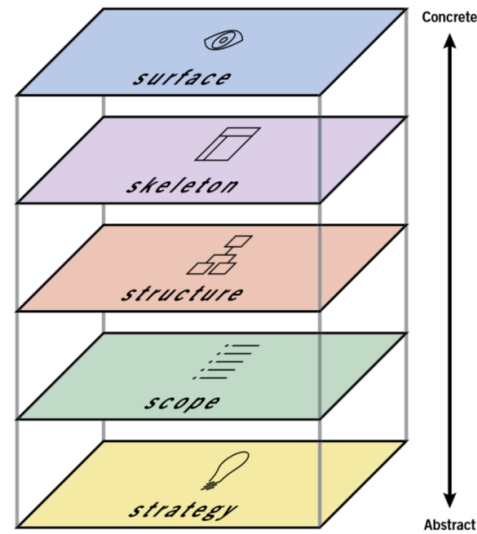
Strategy defines the user needs and business objectives. This is where the objectives of the proposed product are gathered which covers the reason for the product development, who will be the users of this product, what user expects and why they need it. Achieving this goal is done through strategic research process which carries out interviews and brainstorming sessions with users and stakeholders. The scope defines functional and content requirements of the system functional requirements are functions and features the system is supposed to provide to the consumer. The content requirement is the information need to be provided in order to uphold value [3]. Up to now we talked about the strategy and scope these are more abstract elements of user experience. From structure onwards, these elements tend to be more visible without digging deeper in analyzing a product. The structure is a combination of Interaction design and information architecture. Interaction design defines how user interacts with the system and information architecture is how information content is organized. Skeleton layer determines the presentation of the elements and their arrangement. Skeleton layer is based on three components, Interface Design, Navigation Design, and information design. Skeleton would provide the interaction between the elements on the user interface. How the information is presented, how content is arranged and how users are navigated within the application. The most concrete layer Surface is the visual design the appearance of the elements of the application. The look of the design when we are talking about the "look and feel" of a product.

In this research, the focus will be on the top three elements. Surface, Skeleton and the structure. Personas will be used to capture these elements and persist user behavior on different element implementations.

1.3 User Personas

Personas are a representation of users of the product which defines what type of user he/she is.in User Experience engineering personas are used to help the design engineers to align with different strategies and goals of user groups. Engineers apply those personas to their product design and development process. However since personas are based on user attitude, characteristics, behavior and etc. they change over time [24]. Personas can be in different types. The credibility of personas are questionable since these are based on interviews there might be flaws in capturing actual data since interviewee might not disclose their real idea about the system.

1.4 Crowdsourcing user experience engineering

Crowdsourcing is where a set of organizations or individuals contribute in achieving a common goal. The advantage behind crowdsourcing is it creates more value than a set of few users could. We identified that creating personas through a limited set of users and evaluating user interface is not credible. However, crowdsourcing user experience evaluation would provide better results because many users contribute their personas will match to create sets of personas and through their decisions. However using crowdsourcing in user experience evaluation context is a challenge. We will be discussing these challenges and ways to overcome them later in this dissertation.

1.5 Research Area

The purpose of this research is to identify out of those seven factors what can be chosen to evaluate user experience of web applications and based on those evaluations how to improve the user interface of the website that will be desirable for the majority of the people accessing the website while maintaining personal preferences as well. Normally in the software life cycle user interface/user experience engineers develop wireframes and mockups to give an aesthetically pleasing design to the application. Then the client will approve these designs based on their experience and domain knowledge. But in most of the cases, web application developers fail to provide a better design. This is because the users' perspective of the design will change with time. And current surveys and questionnaires fail to provide a better insight to user's attitude about the system, also the task of answering survey questions are less engaging to users that they will skip providing valuable feedback. Capturing this highly dynamic and changing user's attitude towards the system over time is more engaging and consistent when using conversational interfaces. Capturing this user behaviors and ideas is done through creating personas and by questionnaires, questions are invoked from the conversational interface while providing humanly conversational experience.

1.6 Research Question

How can conversational interfaces improve quality of questionnaires for the user experience evaluation, be more engaging and consistent in carrying out humanly conversations in-order to capture users attitude towards the system.

1.7 Research Scope

This research will focus on Identifying the factors that influence in providing better user experience, based on these factors how to capture users' desires, attitude towards the system, How crowdsourcing would help in user experience evaluation, How effective conversational interfaces will be in capturing these from humanly conversations than the traditional questionnaires and what insight would it provide in using personas demographic information along with captured survey data. To research on crowdsourcing eco-system will be difficult. Therefore, this segment is meant to limit the scope of this research in order to make it more controllable. The research is defined by the following boundaries and considerations:

- Case study is conducted within executives of SriLankan Airlines
- Selected users were in the age range of 25-60
- Participants were not provided any guidelines in how to use conversational interfaces

1.8 Motivation

The objective of this research is to capture the changing perspective of the users on the user interface of web applications. The motivation behind this research is to attempt to identify the factors that influence in designing a better user interface and how users react to the design. For instance, this would determine why some designs stand out from others.by crowdsourcing user evaluations would be captured and these data would determine what preferences each persona will prefer. The changed aesthetic design, color schemes, navigational structure, information architecture could be personalized for the user and the web interface would adapt to the majority preference where it will be used for the guest of the web application. A personalized interface would be the ultimate goal of the research however currently it focuses on gathering information on users' attitude towards the system which paves way to personalized interface, through conversational interfaces for usability evaluation questionnaires. And this application is capable of integrating to any web application. This study will also evaluate the impact of using conversation-based questionnaires which are capable of maintaining humanly conversation with system users to capture users' desires, attitude towards the system.

CHAPTER 2

LITERATURE SURVEY

2.1 UX evaluation techniques

Our focus on this research will be on evaluating the user experience of products that are already on the market. Since there is a product that needs to be evaluated the normal procedure would be researchers or the engineers would approach the people who have used this system before and use interviewing techniques to capture data. Because they have been using the system for a while these participants have the true motivation to use the system however we can also evaluate the product with participants who have not used the system before, then it would capture how a new user will be comfortable with handling the system. Through crowdsourcing these participants can be reached without any hassle of approaching them physically, but there should be strong mechanisms to capture user experiences and emotions. Here we will be discussing some user experience evaluation techniques that are mostly suitable for products which have been deployed and currently on the market [4].

Aesthetics scale evaluation method can capture quantitative measures on aesthetic quality of a particular website. This was developed by Lavie and Tractinsky where they conducted four studies in order to develop a measurement instrument of perceived web site aesthetics. Using exploratory and confirmatory factor analyses they found out that users' perceptive about a product usually consist of two main dimensions, which they termed them as ''classical aesthetics'' and ''expressive aesthetics'' [5].

Affect Grid evaluation method was published as a single-item measure of the two affect dimensions pleasure-displeasure and arousal-sleepiness. Used as a way of assessing affect where the participant will mark their emotional state on a 2-dimensional 9x9 grid where y-axis is arousal forms and x-axis is pleasantness.

Evaluation through emocards is at the end of every task or user story completion the participant is asked to pick a card which consists of cartoon faces which describes the feeling or the emotion that user felt about the interaction with the product. Emocards are a quick and simple way to evaluate a product which is cost effective at the same time.

WAMMI (Website Analysis and Measurement Inventory) is a commercial software for evaluating usability which measures website user experience, tracks

changes to the website user experience over time, generate dashboards for the upper management to clearly distinct website usability changes. WAMMI uses about 20 statements in their questionnaire which is focused towards capturing participants personal views towards the product.'WAMMI' button can be placed on master pages where visitors are encouraged to click and provide answers for the questions. These questions are standard and new questions can be added based on client's needs. Based on the gathered data WAMMI will benchmark the website relative to the other websites and the generated report will provide analytics and qualitative comments about how to improve the website usability. However one issue with this technique is that the decisions we take through WAMMI can easily be clouded by providing false information. We can't guarantee the expression the participants provide is reliable. Another issue with WAMMI is that it doesn't capture users' behavioral data.

Researchers have evaluated currently using user experience evaluations and most of the evaluations are based on questionnaires and it has been proven to provide better results. [6] Questionnaires are typically surveys conducted to gain feedback from the perspective of the users. Users provide information about their personal preferences on the design of the user interface and overall experience of using the product [20].Steps of creating a proper survey to capture these data and conducting the survey can be summarized as follows [21],

- Set the goals - What do you want to capture?
- Decide on the target population and sample size - Who will you ask?
- Determine the questions- What will you ask?
- Pre-test the survey - Test the questions
- Conduct the survey - Ask the questions
- Analyze the data collected - Produce the report

Most important step is setting the goal of the questionnaire/Survey, since goals will determine the target user group and what to ask in the questionnaire. Researchers can use three types of questions in their questionnaires such as, multiple choice questions,

numeric open-end questions and text open-end questions. Guidelines to design a proper survey also contains how the questions should be ordered. Questions should be ordered such that easier questions are presented first and harder questions will be presented at the end [20]. There are also ethical guidelines in designing a questionnaire. The American Statistical Association provide Ethical Guidelines for statistical practice they describe the ethical behavior as "The discipline of statistics links the capacity to observe with the ability to gather evidence and make decisions, providing a foundation for building a more informed society. Because society depends on informed judgments supported by statistical methods, all practitioners of statistics, regardless of training and occupation or job title, have an obligation to work in a professional, competent, and ethical manner and to discourage any type of professional and scientific misconduct" [22]. Let's look at some questionnaires used in user experience evaluation and how they capture user behaviors based on the well-designed questions.

IsoMetrics is an inventory developed by Gediga, Hamborg and Willumeit [8] for testing of software usability. IsoMetrics is based on ISO 9241/10 standard. This standard is an internationally recognized comprehensive set of rules for ergonomic requirements of computer-based office work which intended to achieve usability of software systems. Part 10 of the standard is focused on organization and assessment of dialog systems. Suitability for the task, self-descriptiveness, controllability, conformity with user expectations, error tolerance, and suitability for individualization and suitability for learning are key design objectives discussed on this standard. IsoMetrics questionnaire is catered for both *summative evaluation* and *formative evaluation*. Both terms summative and formative are from educational theory where formative being providing immediate feedback to improve learning and summative being evaluation of what was just learned .formative usability evaluation is fixing problems of the interactive design process and summative evaluation focus on the current usability of the user interface, which is measured through completion rates and task times.

		strongly disagree	disagree	so - so	agree	strongly agree	
Index	**Suitability for the task**	**1**	**2**	**3**	**4**	**5**	No opinion
A.1	The software forces me to apply procedures and/or routines that are not related to my actual work.						
A.2	I have the feeling that the software requires me to carry out actions that really have nothing to do with the task I am trying to perform.						
A.3	The software lets me complete entire work routines.						
A.4	The built-in functions of the software support me in completing my tasks.						

Figure 3: IsoMetrics questionnaire

Figure 3 shows example of IsoMetrics questionnaire which is used for formative usability evaluation. IsoMetrics can be used for both individual and group settings. Though the evaluation of the results when subject ticks "no-opinion" over at least 20% of all the queries presented these answers are eliminated without further study.

SUMI (Software Usability Inventory) was developed by Kirakowski and Corbett for the project "Metrics for usability standards in computing". SUMI is a solution for recurring problem of users' perspective on the usability of the software. It provides an objective way to assess users' satisfaction of the software. SUMI consists of 50-item paper-based questionnaire. Where users responses are mapped on a three point scale as agree, undecided or disagree. Following are some sample questions which were in SUMI [10],

- This software responds too slowly to inputs.
- I would recommend this software to my colleagues.
- The instructions and prompts are helpful.
- Learning to operate this software initially is full of problems.
- Working with this software is satisfying.
- The way that system information is presented is clear and understandable.
- I feel safer if I use only a few familiar functions.
- I think this software is consistent.

SUMI questionnaire is very effective to assess new products during evaluations, to make comparisons on different versions of the product and to set usability targets for future product that are willing to be developed. Analysis of SUMI data is done through SUMISCO software. SUMISCO gives diagnostic information of queries and identifies whether user responses are negative or positive. Which can be used to identify that users have difficulty in a particular task. SUMI is multilingual, supporting each language translated and validated by native speakers. SUMI captures user satisfaction based on five aspects, Efficiency, Affect, Helpfulness, Control and learnability. Efficiency is that the feeling user will have that he/she can complete task in a shorter time period. Affect is the psychological emotional feeling that the user has towards the system. It refers to that user feels mentally stimulated and pleasant. Helpfulness is that the interaction with the software user is helpful and assists in a way that user can easily perform his/her tasks Control is that user feels they have control over the product and Learnability is how easy is the system that a new user can get started quickly and come up to pace of an experienced user. SUMISCO will generate reports based on these aspects with the data captured through the questionnaire.

Laugwitz, Held and Schepp's evaluates of user experience based on another set of factors Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty [11].In order to capture these factors they develop a questionnaire. For their work they relied on a theoretical framework of user experience which distinguishes ergonomic quality, hedonic quality and attractiveness of the product. Ergonomic quality emphasizes that user can perform their task achieve their goals efficiently, while hedonic quality focus on non-task oriented aspects of the product such as aesthetic pleasantness of the design/user interface. With their study they identified that the questionnaire should contain two classes of items. Namely, items which measures the perceived attractiveness directly and items which measure quality of the product. To generate the terms for the questionnaire they invited group of experts to propose terms for the assessment of user experience.

The experts were given following questions to answer,

- To which properties of products are users particularly responsive?
- Which feelings or attitudes of users are caused by products?
- What are the typical reactions of users during or after usability studies?

229 expert answer proposals were submitted once redundant queries were removed. Then seven usability experts extracted top 25 sets of terms. Removing inappropriate adjectives 80 adjectives remained. The best fitting antonym for each of the 80 adjectives was identified [11].They use seven stage scale to reduce the well-known central tendency bias. Then the split questionnaire item set into two parts as they described earlier as terms that represent emotional behavior that they don't provide any reasoning for acceptance or rejection and the others. First set of terms were mapped into the factor called Attractiveness though factor analysis and they determined second set of factors according to the items that showed the highest factor loadings as Perspicuity (examples for items: easy to learn, easy to understand), Dependability (predictable, secure), Efficiency (fast, organized), Novelty (creative, innovative) and Stimulation (exciting, interesting).The final questionnaire contains six items from the scales Attractiveness and four items each from Perspicuity, Dependability, Efficiency, Novelty and Stimulation. They called this questionnaire User Experience Questionnaire (UEQ).The generated questionnaire was subjected to a validation study which produced promising results. The proposed user experience questionnaire (UEQ) is an easy to apply, reliable and valid way of measuring or evaluating user experience.

This idea of constructing User Experience Questionnaire (UEQ) was again revisited by Rauschenberg, Thomaschewski, Cota and Schrepp where they clearly describe the dependency between the evaluation scales. This questionnaire format supports to capture user's response based on their attitude towards the product, impressions, and feelings [12]
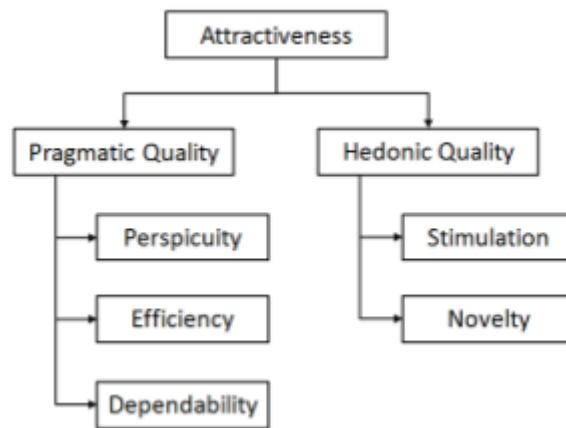
Figure 4: Dependencies between aspects of usability

Manzoor [13] has come up with a usability tool which is designed for higher educational websites and evaluated the tool using a questionnaire. The usability evaluation tool consists of broken link checker, page availability test, checking links on words. Checking image sources without "ALT" attributes, checking website performance and statistical analysis. Broken link check will go through the links of the website and point out where the links can't be redirected to a specific path. Results would be given as how many broken links were found. Page availability feature will search for sitemap page, contact page, about us page, home page and etc. Checking links on words module will check for words such as more, click here and links which decreases the usability of the website. It'll also check for nested links and provide the result in total number of links found on each section. Checking for image sources without "ALT" attributes is for checking the website is designed visually impaired people on mind which would help them to determine what the image is about. Website performance check would provide data about the load time of the website excellent performance is where website loads in less than a second. Good performance is greater than 1 second and less than two seconds. Average performance if it's greater than two seconds and less than three seconds and finally poor performance if it's greater than three seconds. However this assumption is quite vague since it mostly depends on the dynamic content that is provided from the website and differs with the context. Statistical analysis was carried out to determine the overall result of usability evaluation.

Layla Hasan [14] has provided a usability evaluation of educational websites based on students preferences of design characteristics. The usability evaluation mechanisms are categorized into three sections here as user-based methods, Evaluator based methods and tool based methods. User-based methods involve the user identifying usability problems by participating in questionnaire or interviews Evaluator base method is identifying usability problems through heuristic evaluation and tool-based evaluation is usage of software tools to analyses the HTML content using many guidelines. The motivation of the study was through increasing the usability of the educational websites uphold the overall learning experience, build proper confidence in students using the website and encouraging them to refer the sources more often. Criteria for evaluating website is developed which consists of five main categories, Navigation, Architecture/Organization, Ease of use and communication, design and content. Navigation category consists of five sub-categories navigation support, internal search, working links, no broken links and no orphan links. Architecture/Organization consisted of three subcategories logical structure of the site, not deep structure and simple navigation menu. Ease of use and communication consist of four subcategories quick downloading of web pages, easy interaction with website, Contact us information and foreign language support. Design category would focus on Aesthetic design, appropriate use of images, fonts and colors, appropriate page design and consistency. Content category assesses whether the website comprises of rich information based on their context. Subcategories of focus are up-to-date information, relevant information, No under construction pages, accurate information, information about the university, information about the colleges and information about departments. Students were instructed to assign weights to each category. According to the results, the usability categories were ranked according to the descending order as navigation, content, ease of use and communication, design, and organization/architecture. Interestingly when they evaluated results based on gender. Female participant has rated content over navigation. This research provided insight on the relative importance of design categories from the viewpoint of students unlike the earlier researches carried out. This research also investigated the gender impact of the relative importance of usability criteria. However, the study was limited to students

where it left out stakeholders of educational websites and the research employed only user-based evaluation.

UX curve method is a retrospective method for evaluating user experience over time. Users sketches on a curve how his/her relationship with the product changed overtime based on his/her opinion [15, 16].The template is a two-dimensional graph, it starts from the first impression of the product to the present in the horizontal axis and vertical axis represents the intensity of user experience with a positive upper part and a negative lower sections marked with "+" and "-".User marks the peaks and lows explaining what caused the change in the lifetime of the product. Curves can be drawn focusing on different aspects of user experience such as attractiveness, ease of use, degree of usage and etc.
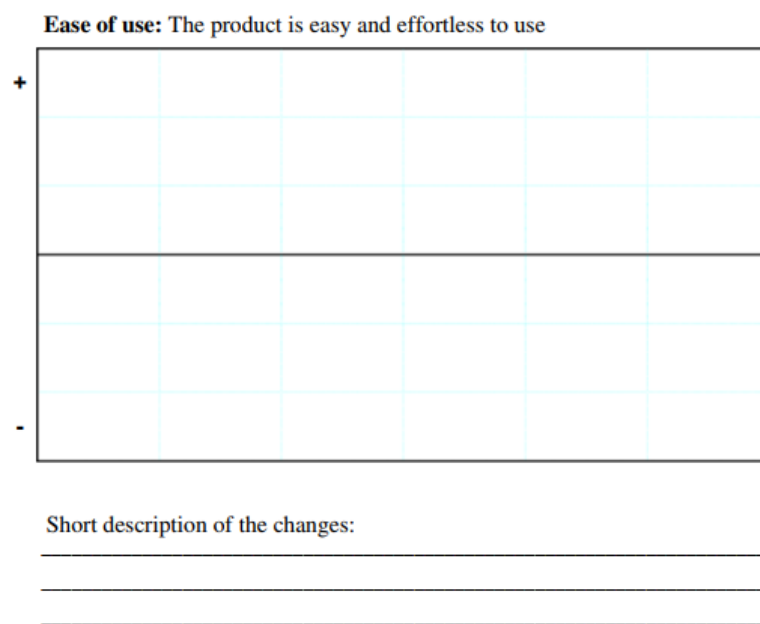
Figure 5: UX curve

UX curve method is validated using two questionnaires evaluating user satisfaction and customer loyalty. However the UX curve method relies on users' memories about their experiences the time period varies from 1 week to 16 months. So these responses they provide might not be exact experiences they had in mind at the time. Though the study they also pointed out that attractiveness curve seems the most powerful curve

covering broadest view of user experiences revealing pragmatic and hedonic perspectives of the system [16].

2.2 Crowdsourcing techniques

User-based evaluation is important technique to capture user experience, user's attitude towards the product to enhance user satisfaction. These techniques vary from simple surveys to laboratory studies. This collection of input from the participant are mainly carried out through questionnaires. However these evaluations are limited by a small set of participants. In testing the usability of a product it is desirable to evaluate the product with many participants that is possible since it would increase the possibility of discovering issues and errors in the product. This lead researchers to focus on online surveys where they are capable of reaching out as many participants possible. Crowdsourcing is a widely known online production model capable of providing vast success in requesting work and completing work which involved a transformative effect on many disciplines as well as in human computer interaction [27]

Amazon Mechanical Turk [17] is a micro-task oriented workplace where requesters submit micro-tasks or work that will be completed by the workers online. The workers will be granted a monetary reward. Using Amazon Mechanical Turk for user studies is a challenge since the system usually shower higher success rate in micro tasks however user experience evaluation employ complex tasks as well. There is also the challenge of guaranteeing authenticity of answers provided by the users, because workers are provided with many tasks in order to complete many tasks per hour to gain more money they would be tempted to provide nonsense answers that won't add any value to the evaluation. Another challenge is lack of demographic information, unknown expertise of the user and limited communication with the participant where is user experience evaluation target groups are drawn around the globe with known expertise and can be interviewed for more information. The researchers carried out two experiments to analyze how Mturk would be behave on usability evaluation. The task was to do a quantitative evaluation based on user ratings and feedback of

Wikipedia articles. Though the first experiment they observed there is a 58.6% potentially invalid responses from minority of users. For the second experiment they altered the method of collecting user experiences that they match it with expert responses. Users were instructed to provide input for four questions with quantitative answers accompanied with a rating of the article. Second experiment provided better results with only 7 meaningless responses from turkers. Mturk platform would provide promising results if the tasks were designed properly. Input from hundreds of users can be gathered in larger geographic area.

CrowdFlower is another service that provide crowdsourcing "channels" same as Amazon Mturk. The potential of these networks were studied by Liu, Lease, Kuipers and Bias [18] CrowdFlower allow requesters to upload their task to Mturk to be done by the workers or any other crowdsourcing channel that requester desire. CrowdFlower will break down these heavy projects to more manageable micro-tasks that will eventually completed by the workers. And the result of each task would be aggregated to produce a final result. To reduce "spammers" or users providing nonsense answers they have specific quality testers known as "Gold Units" these are hidden tasks random distributed across the main task and they have known answers if a "spammer" provide wrong answers to the question his/her answers will be automatically rejected. uTest is another crowdsourcing company which is specialized towards usability testing. The advantage of uTest over Mturk is in uTest customers can specify tests to various user groups based on geolocation, platform and devices which users are working on, while Mturk users remain anonymous. In their study [18] they conducted two usability tests on a graduate school's website, one using traditional lab usability test and the other using crowdsourcing. For crowdsourcing the used Mturk platform and using CrowdFlower as an intermediary. Both types of user found out similar design flaws of the system however they identified that in the lab study if the question is not clear the participants were capable of asking questions and in crowd sourced study workers couldn't request any help and none of them even bothered. So it is important how the instructions are provided for crowd sourced usability tests. Instruction should be clear, more descriptive and unambiguous.

Intelligent user experience questionnaire (IUEQ) is an intelligent questionnaire generator for the participants. Main concern was rather than asking the participants directly "Do you like or dislike this software ?" the user's idea would be about the overall software however with IUEQ we can generate questions based on GUI elements and by considering user's context. Job creators were provided with a template where they can create rules based on attractiveness, efficiency, perspicuity, simulation and novelty. Rules were capable of capturing user factors, social factors, context of use and product factors. These user input were cross-referenced with user details provided at registration to provide a score for trustworthiness. Their main focus of usability study was mobile devices where they integrated an avatar to the android application.

We discussed how crowdsourcing can help to enhance user experience evaluations, we also discussed the challenge of removing spam responses created by the users. It is identified that processing of crowd sourced information should be carried out to maximize the reliable, useful information.txteagle is a system that enables mobile users to complete simple tasks and gain a monetary rewards which was launched in Kenya. Tasks include translation, transcription and surveys. In order to validate the information provided by the users they send out the same task repeatedly over to many users and verify that we are getting the same response. They assess the validity of a response through majority voting. They infer the most popular response from set of users and the probability of response being correct is noted. This majority voting scheme is adequate for simple task completion but to access the accuracy of individual user, maximum likelihood estimation is carried out by expectation maximization (EM) model [30]. Mechanical Turk also uses this technique to filter out noisy responses from error-prone individuals or scammers. Another approach is Reputation and trust modelling. Trust evaluation in crowdsourcing systems are quite complex because workers are anonymous in most of crowdsourcing systems (e.g. Amazon Mturk), interaction between the worker and work requester is low and the diversity of the completed tasks. It is observed that on different types of tasks and reward amounts the workers tend to show different trust levels. CrowdTrust [31] is trust modeling method model trustworthy worker selection problem as a multi-objective combinatorial

optimization problem without subjective weights. So it doesn't assign weights for trust based on type of the task or the reward associated with the task. They have used NSGA_II algorithm for solving the multi-objective optimization problem. Another challenge in crowdsourcing is herding. This is a result of reputation based task evaluation we used to differentiate spammers from genuine workers. Essence of crowdsourcing in user experience lies at the result of the usability evaluation is from different users from different knowledge levels, contexts and accessing content on different devices. Imagine a situation where workers are elected based on their reputation, workers from highest reputation might be from the same background. Herding refers to the situation where task requests consist with only a small group of reputable workers causing them to overload while leaving other workers idle [32]. It can lead to system failures. Herding is discussed with system failure to cater high traffic to the system however imagine the questionnaire being overloaded with same set of answers.

2.3 Persona creation based on questionnaires

Personas are fictional characters that user experience engineers create to understand the user and his/her needs. In User Experience engineering personas are used to help the design engineers to align with different strategies and goals of user groups. It's identified that these personas change overtime time since personas are mostly used in the design phase of software development sometimes these changes of personas will never be evaluated again in the life cycle. According to a research done by Norman group [23]. This survey was carried out by participation of 156 user experience professionals and nearly 46% update their personas every 1-4 years.28% update personas quarterly or more often and another 26% have not done revision after 5 years or not at all. The survey has also rated how successful the personas were in upholding the user experience. Average persona impact ratings were higher in people who had frequently revisited their personas.
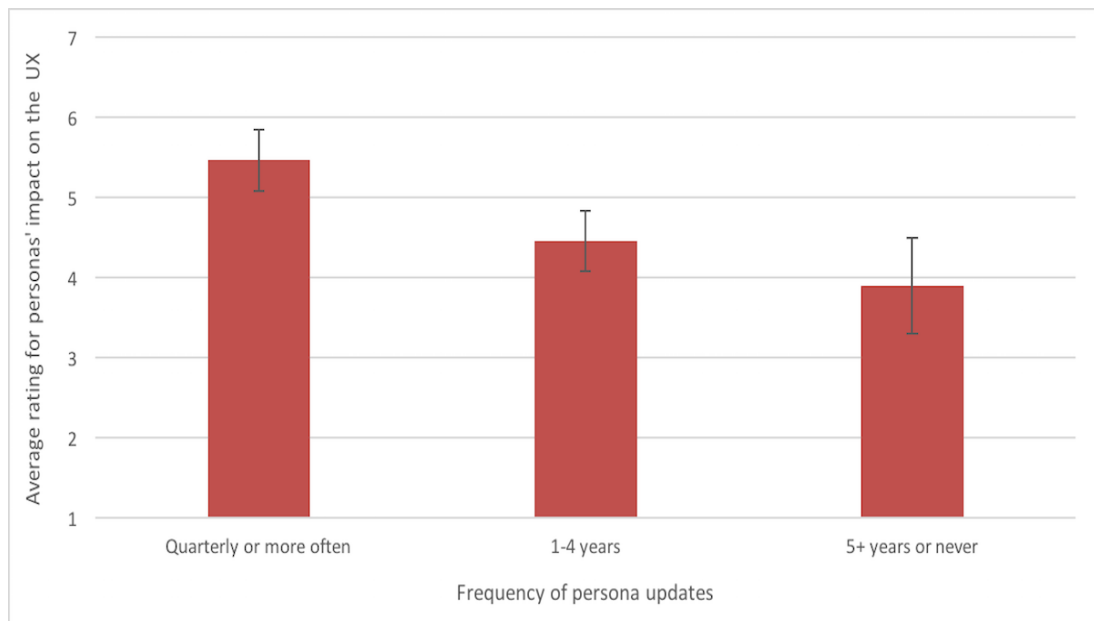
Figure 6: Average personas' impact rating segmented by revision frequency. The bars represent 95% confidence intervals [23]

Personas have a major impact on user experience, after-all they are real user groups who get affected by the product. Monitoring changes in personas and reevaluating them will express the real needs and expectations of target user groups [24].

In the Persona project at Microsoft Research [25], they have developed a prototype that employ and explicitly anthropomorphic character that interact with the user in the natural language. It's identified that users will be unwilling to speak to the computer in command languages when the interaction between the computer and the user is more conversational they responded positively. However the assistant should be capable of understanding broad range of English paraphrases of user's intent. It is studies that the above interaction inevitably evokes human social responses that users are more confident in providing information to the life like character that they tend to treat computer as human [26]. Involving these in user experience engineering for persona creation we will be using these assistants to carry out interviewing process. In IUEQ [19] they have also use this approach of an avatar to interview the participants. This approach should also be further evaluated since in the context of usability evaluation and transforming the questionnaires to interactive conversations should be further studied to provide clear understanding that how user would react to such scenarios.

2.4 Conversational Interfaces

Conversational interfaces are capable of mimicking a humanly conversation. These interfaces are infused with a persona to give a personality to the interfaces such that these interfaces will have own likes and dislikes providing them to be more human rather than using graphical user interfaces to perform tasks users can communicate with the interfaces to get their work done. Chat bots and voice assistants have emerged to businesses in order to provide services to the customers. These conversational interfaces use Natural Language Processing in order to understand what users try to communicate and map them with the exact command which performs the task as in the graphical user interface. Natural Language Processing (NLP) structure the human language such that where an utterance is provided, it performs the identification of intent or the actual task the user prefers to perform. NLP system should be capable of ingest an utterance of the user, comprehend the meaning, perform the preferred task and respond to the user or provide reliable information to the user. Natural Language Understanding (NLU) is a subset of NLP. NLU is the post processing of the parts of speech identified by NLP. Chatbots are web or mobile based conversational interfaces which are used to market products and enable ordering through the conversation interface or providing information to the users acting as a customer support agent.

Personalizing Dialogue Agents [34] research focus on assigning a personality to the agent so that chit-chat model dialogues are more engaging and consistent. The research team has identified problems in chat-chat models. That they doesn't display a consistent personality. This is due to the fact that these models are trained with dialogues of different speakers and not categorizing them on the persona of the speaker. Due to the fact that these systems are trained over short term utterances which only capture recent dialogue history they are incapable of having a long term memory. And the tendency of providing non-specific answers to users utterances, these problems combined will reduce the overall experience that users are discouraged to engage with the conversational interfaces. The researchers have identified these problems mainly arise due to the fact that there are no good publically available

datasets for general conversational dialogues. Their research aim to provide more engaging chit-chat dialogue agents with a persistent persona. Where this persona is stored in a memory augmented neural network and provide consistent responses to the users emphasizing the personality of the agent to provide more engaging conversation. They have introduced a PERSONA-CHAT dataset which contains chit-chat dialogues over 162,064 utterances. These were captured using Amazon Mechanical Turk (MTurk). Where each pair of crowd workers were asked to act the part of a given persona. They were asked to chat naturally and get to know each other in a conversation. The data collection for PERSONA-CHAT consist of three stages. Namely, creation of personas, revising personas and initiate persona chat. They have crowd soured 1155 possible personas in this study. For revising personas they have re-written the exact same personas with related sentences that are rephrases.

Original Persona	Revised Persona
I love the beach.	To me, there is nothing like a day at the seashore.
My dad has a car dealership	My father sales vehicles for a living.
I just got my nails done	I love to pamper myself on a regular basis.
I am on a diet now	I need to lose weight.
Horses are my favorite animal.	I am into equestrian sports.
I play a lot of fantasy videogames.	RPGs are my favorite genre.
I have a computer science degree.	I also went to school to work with technology.
My mother is a medical doctor	The woman who gave birth to me is a physician.
I am very shy.	I am not a social person.
I like to build model spaceships.	I enjoy working with my hands.

Figure 7: Revised versions of original persona [34]

In the revision task crowd workers were instructed not to rephrase the sentence using the exact words. After collecting these dialogues crowd workers were paired randomly and were instructed to chit-chat with each other while playing the persona they were assigned.

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Figure 8: Captured conversation between the personas [34]

With the collected chit-chat dialogues they have performed evaluation of the dataset by testing various baseline models on the dataset. Through these evaluations they have identified that models that have access to the personas were scored as more consistent by the annotators, however not more engaging the models trained on PERSONA-CHAT were more engaging than the models trained on other dialogue sources. These collected data will be useful in creating of agents that are having consistent personalities and perspectives. This research took and important step in modeling dialogue agents that are capable of asking personality based questions, remember the answers provided and use these in natural conversation interfaces. This research have highlighted the need to have a personality with likes and interests to capture customer's attitude, feelings towards the system. However these conversations were not naturally captured and sometimes users tend to have more complex conversations with each other and these conversations should be performed in vast number of varying personas.

Pierre- Emmanuel, Samuel Humeau, Martin Raison and Antoine Bordes revisited the Personalizing Dialogue Agents research paper [34] and second their idea of having personas with personalities increase the engagement of the end-to-end conversation level. However they emphasize the conversations were synthetic and the variety of personas taken into the study was limited. In their research they try to broaden these

limits introducing a net dataset providing 5 million personas and 700 million conversations based on the personas [35]. They aim to show that even at this scale training using personas still improves the performance of end to end systems. In order for dataset creation they used utterances of REDDIT comments, and they organized them as Persona, Context and Response. Persona is the sentences representing the personality of the utterances. This contains likes and dislikes of the user. The context is the utterance which it should respond to and the response is the predicted answer by the system. Their goal is to predict responses based on natural conversations. Through identified Persona and Context they encode context and persona which will be combined using a 1-hop memory network with a residual connection. All the candidate responses and the representation of context and persona were inputs in obtaining the dot product. All the dot products were passed through softmax in order to maximize log-likelihood of correct response.
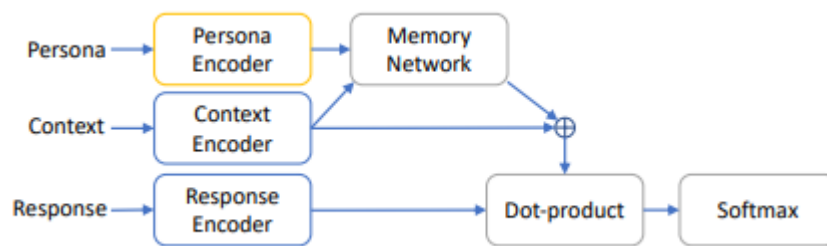
Figure 9: Persona based neural network architecture [35]

In sample predictions from the model personas consists with a single sentence and predicted answer contained at most 10 tokens which have been retrieved from responses from candidate samples collected from the training set.

Context (Persona)	Predicted Answer
Where do you come from?	
(I was born in London.)	I'm from London, studying in Scotland.
(I was born in New York.)	I'm from New York.
What do you do?	
(I am a doctor.)	I am a sleep and respiratory therapist.
(I am an engineer.)	I am a software developer.

Figure 10 : Question context and predicted answer [35]

The core concept of the research is modeling dialogue by next utterance retrieval and the response is selected through set of candidate answers. The team demonstrated that training using personas will improve the performance in chit-chat conversation based systems.

CHAPTER 3

METHODOLOGY

The aim of this chapter is to provide the pathway through which objects of this study is achieved. It presents the logical assumptions behind this research, as well as introduces the research strategy and development methodology was observed techniques applied.

3.1 Software Development Methodology

Throughout the research project, software development life cycle is based on Iterative model. Iterative model focus on an initial simple implementation of the system which will be iteratively or progressively improved with complexity and a wider range of features until the final system is completed. Iterative model is a cyclical process where after the initial planning phase. Each phase is repeated over and over with each completion of the cycle improving and adding new functionality. This research focus on developing a Chatbot which is capable of carrying out the task of User Experience evaluation questionnaires. In order for Chatbot to be trained over a number of Intents. Questions and Answers were designed iteratively where first set of questions were configured and they were evaluated whether they can capture user's intent successfully. Also the features such as Sentiment Analysis are to be added in later iterations replacing the score scales of the questionnaires.

3.2 System Design

The intermediate application is developed to capture questions invoked by the UI/UX evaluation Chatbot, Answers which users provide, Sentiment score of the provided answer. The intermediate application is based on a three-tiered architecture. Intermediate application will provide the functionality to maintain the dashboard which provides insight on answers provided by the users and their sentiment scores, it's also capable of configuring questions to be asked by the Chatbot and question contexts.

3.3 Approach and Technology Selection

3.3.1   Interviewing participants and invoking questions

Though the research on creating personas from interviewing the participants we discussed that the avatar based interviewing solution would be more desirable over traditional methods.to achieve this assistant should be capable of understanding broad range of English paraphrases of user's intent. We will be using bots to maintain the conversation with the user so user will feel that he/she is having a conversation with another human being.so in order to provide this language understanding features to the bot Language Understanding Intelligent Service (LUIS) will be a candidate option which is a part of Microsoft cognitive services [28]. In order to understand a query from the user entities of the query to be understood by the system. There are three concepts associated with it namely, utterance, intent and entities. Utterance is the query from the user, intents are the actions that user want to follow or user's intentions and entities are the real world objects in the query.

Table 3.i : Utterance, Intent and Entity

Utterance	Intent	Entities
What is your name?	To know name	You
What is your age?	To know age	You

LUIS [28] should be properly trained by providing utterances, intent and entities in the beginning after our endpoint has processed some interaction LUIS will begin active learning that the system is capable of understanding the utterances and calling required functions. Figure 11 shows a Chatbot which is used to provide information to the customers or support in certain tasks in commercial websites the interview for persona creation would also follow the same design where Bot would maintain a conversation between the participants until it gathered the required information to build the persona.
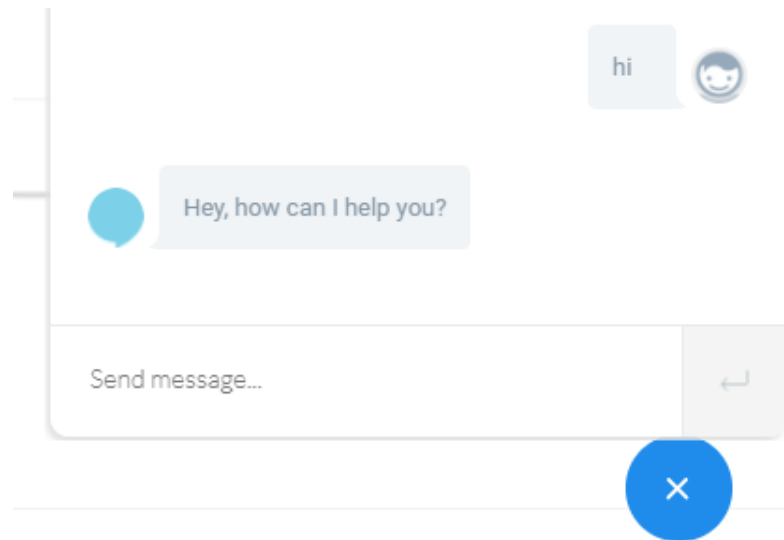
Another preferred solution for this language understanding functionality is Dialogflow [33]. Dialogflow users Natural Language Processing (NLP) backed my Machine Learning. This process runs on Google cloud which provides an infrastructure which is capable to cater requests from millions of users. While dialogflow flows the same structure which is the basic structure of a conversation. Normally traditional interfaces require a structured input in-order to process information. However in a humanly conversation inputs are not structured nor predictable and context of the information vary rapidly. Natural language processing provides a solution to this problem where it converts the unstructured input of the user to a structured input by understanding what user intends to perform and match it with the relevant function and execute the action needs to be taken.

Consider a sample question where user need to know the current weather. To cater this request user can ask many questions, such as

- What is the weather right now?
- Tell me about the weather in Colombo?
- What will be the weather forecast tomorrow?

Even these are simple questions as seemed by the humanly conversations, our interfaces should be capable of identifying the Intent of the user which is knowing the weather. However in-order to provide a correct answer we need to identify a set of parameters and Entities associated with the request these can be crucial in conversation where follow up questions need to be asked in order to clarify the exact question of these should be obtained from the conversation context. If we further analyze first question user's intent can be identified as to know the weather and the interface should also figure out at what time users is referring to, text should be analyzed to identify "right now" and "tomorrow" as the time, "Colombo" as a place. Dialogflow language parser is capable of understanding the nuances of the language.

Dialogflow consists of agents, an agent is a module with natural language understanding enabled. An agent can be trained with expected user inputs and it maps to a structured data where it can be identified by the interface. These inputs are defined as intents which is the actual intention of the user which will map user input to actions and responses. For each and every intent we should define possible user utterances which will trigger the matching intent and data that can be extracted from the utterances.
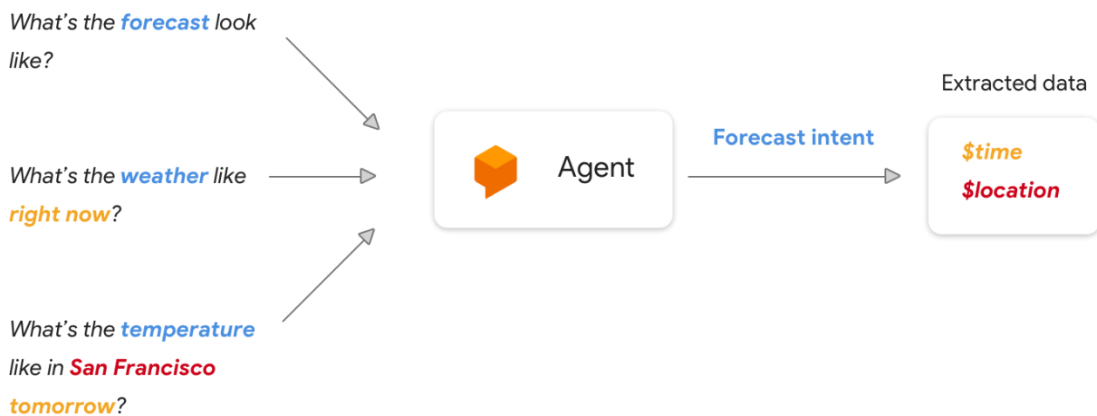

Figure 12: Configuring agent's intent with utterances and parameters

Configuring an intent you need to define Training phrases which will be the input by the user. We don't have to provide all possible user inputs since dialogflow will match the similar utterances and expand the training phrases using machine learning. These

training phases can be annotated with entities, which are data that we need to extract from user utterances. From our former example "Tell me about the weather in Colombo?" Colombo is a parameter value for a location. Now we are capable of matching user utterances with an intent and extract parameters from those utterances then we should provide responses which may provide user with answers, request for more information or terminate the conversation.
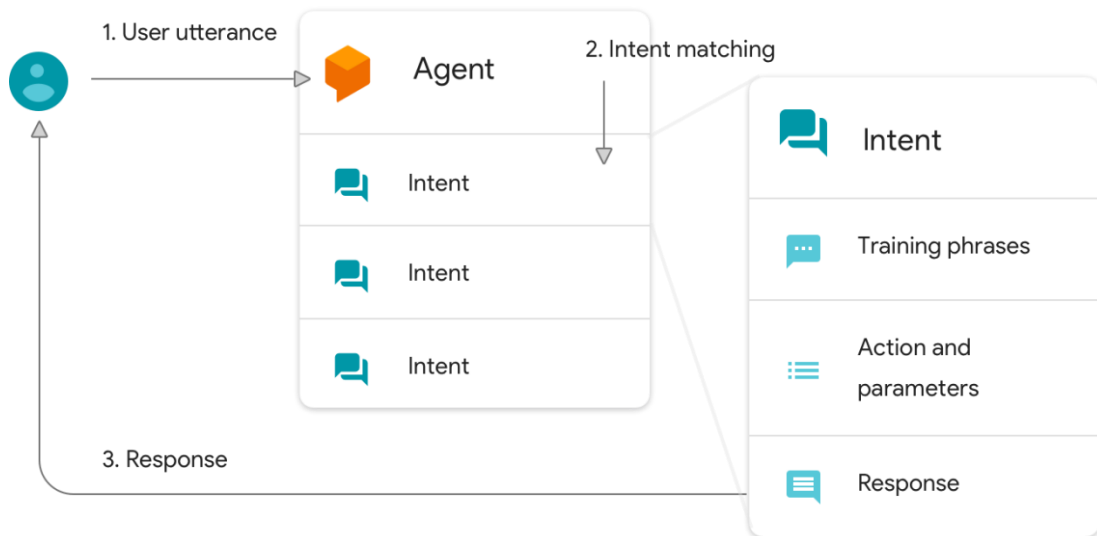
Figure 13 : Configuration of an Intent

3.3.2 Crowdsourcing the user experience evaluation.

Through the literature research we identified that there are many aspects that affect user satisfaction of a design such as efficiency, affect, helpfulness, control and learnability. Questionnaire format should be able to capture user impressions, feelings and attitudes. It should be evaluated whether carrying out these questionnaires through the conversation that we discussed above is desirable or not. It should be evaluated using results from traditional questionnaire and conversation based questionnaire.

Figure 14: WAMMI questionnaire

Above questionnaire is WAMMI questionnaire [29] which consists of 20 statements which has been tested for measuring website user experience. The proposed questionnaire would contain user satisfaction capturing questions as well as personalization based questions that we discussed in the introduction. These questions are used to get users perspective on aesthetic design, information and navigational structure of the website which are defined by top three elements of user experience. Namely Surface, Skeleton and the Structure.

In-order to generate questions for the Chatbot several usability and user evaluation questionnaires were examined. Questionnaires examined were having different characteristics, they were developed over time and started out with large pool of questions. And they were also having different rating scales. Some questionnaire have seven point rating scale while others have three or five point rating scale. Questionnaire normally contain a statement and users were asked to rate their agreements to the statement however when Chatbot asks a question from the user, he/she might provide different answers to it. So questions should not be vague. Questions should be capable of capturing answer for the exact usability factor.

Acronym	Questionnaire	Institution	No. of Questions
USE	USE Questionnaire	Sapient	30 questions
WAMMI	WAMMI Questionnaire		20 questions
SUMI	SUMI Questionnaire		50 questions
PUTQ	Purdue Usability Testing Questionnaire	Purdue	100 questions
CSUQ	Computer Systems Usability Questionnaire	IBM	19 questions
QUIS	Questionnaire for User Interface Satisfaction	Maryland	27 questions
PUEU	Perceived Usefulness and Ease of use	IBM	12 questions
SUS	System Usability Scale		10 questions

Most of the questions were associated with rating scales. However, Chatbot should ask open ended questions in-order to capture the user's real attitude, feeling towards the system. So we need to reconstruct these questions accordingly.

Table 3.iii : Re-phrase questions for chatbot

Questionnaire question	Chatbot question
I am satisfied with it	How satisfied or dissatisfied are you with the system?
It is easy to learn	Do you think it's easy to learn
It helps me to be more productive	Does it helps you to be more productive?

 When you analyze the questions it seems the questions are re-phased In order to capture user feedbacks these question do not have a personality infused with them. If these type of questions were added it would be less effective, less engaging and inconsistent. "How satisfied or dissatisfied are you with the system?" can be made

humanly if this question is transformed to "How satisfied are you with me?" Slight change of the question changes whole conversation environment.

3.3.3 Sentiment Analysis

Sentiment analysis the process of identifying whether a statement is positive, negative or neutral. Sentiment polarity is the main metric of the sentiment. Cloud services like Google cloud Natural Language API, IBM Watson NLU, Amazon Comprehend API, Microsoft Text Analytics and Algorithmia Sentiment Analysis provide sentiment analysis as a service. Most of these services provide a sentiment score as a range between negative and positive ([-1, 1] or [0, 1])

Azure cognitive services provides a cloud-based API for Text Analytics which provides Language Detection, Sentiment Analysis and Key Phrase Extraction, It is capable of processing the raw text provided and identify clues about positive or negative sentiment. It provides a sentiment score ranging from 0 to 1 for each document provided. Here 1 is the most positive while 0 being the most negative [40].

Figure 15 shows how sentiment is analyzed from text using Microsoft Text Analytics API. A document which have a neutral sentiment will have a score of 0.5
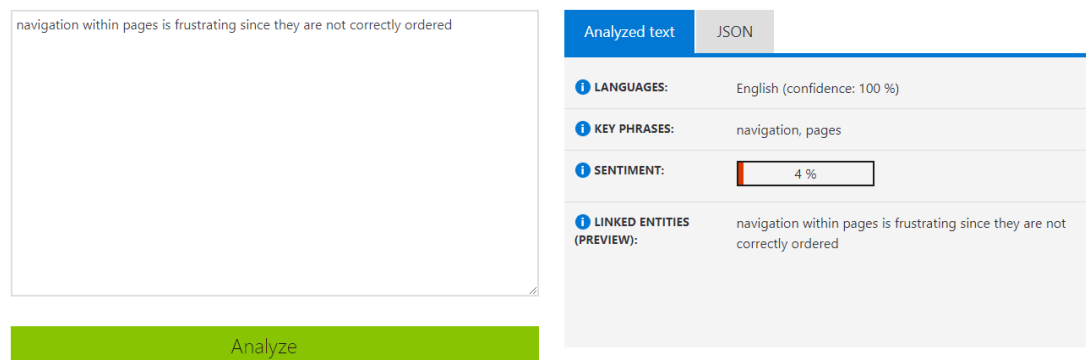
Figure 15: Sentiment Analysis on Microsoft Text Analytics

Microsoft Text Analytics API was selected having expansion in mind. The focus of the research is to provide a generalized questionnaire as a Chatbot to carry out user

experience evaluation through a humanly conversation rather than using likert scales to quantify the user's attitude towards the system. In order for the Chatbot to be integrated to any website we should consider about localization.

Language Name	Language Tag	Speech-to-Text	Text-to-Speech	Telephony	Knowledge Connectors	Sentiment Analysis
Chinese						
↳ Cantonese	zh-HK	✓				
↳ Simplified	zh-CN	✓				✓
↳ Traditional	zh-TW	✓				✓
Danish	da	✓	✓			
Dutch	nl	✓	✓			
English	en	✓	✓	✓	✓	✓
↳ Australian locale	en-AU	✓	✓		✓	
↳ Canadian locale	en-CA	✓	✓		✓	
↳ Great Britain locale	en-GB	✓	✓		✓	
↳ Indian locale	en-IN	✓	✓		✓	
↳ US locale	en-US (equivalent to en)	✓	✓	✓	✓	✓
French	fr	✓	✓			✓
↳ Canadian locale	fr-CA	✓	✓			
↳ France locale	fr-FR (equivalent to fr)	✓	✓			✓
German	de	✓	✓			✓
Hindi	hi	✓				
Indonesian	id	✓				
Italian	it	✓	✓			✓
Japanese	ja	✓	✓			✓
Korean	ko	✓	✓			✓
Norwegian	no	✓				
Polish	pl	✓	✓			
Portuguese	pt	✓	✓			✓
↳ Brazilian	pt-BR	✓	✓			
Russian	ru	✓	✓			
Spanish	es	✓	✓			✓
↳ Latin America locale	es-419	✓				
↳ Spain locale	es-ES (equivalent to es)	✓	✓			✓
Swedish	sv	✓	✓			
Thai	th	✓				
Turkish	tr	✓				
Ukranian	uk	✓	✓			

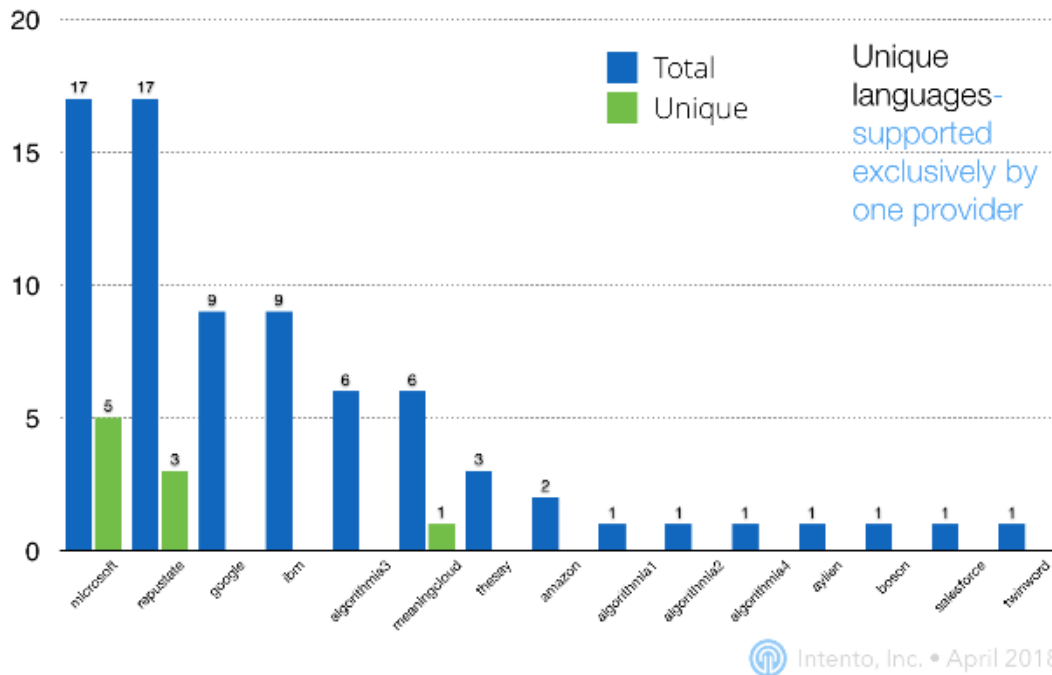*Figure 16 : Languages supported by Dialogflow*

Figure 17 : Language support on Sentiment Analysis services

Azure Text Analytics was selected due to the language support in order for the Chatbot to be enhanced it should possess multilingual support.

3.3.4 Personalization and adaptation

Personalization will be handled from the browser cookies and this adaptation will only be available analyzing data gathered through this research. In order to maintain fewer number of questions, single question is added to capture preferable theme of the user. These data based on data gathered though questionnaires can later be used in user interface design.

For the personalization and customization first designer should be aware of what are the changes that users need, based their view of the system. These information can easily be captured using the proposed solution. We just need to include a question set which is specifically targeting on the look and feel of the website. Imagine a situation where user is an elderly citizen and he/she prefers a larger font size on their information

content. We could include a question to our question set that asks the question "would you find current text size to your liking?", If the user provides a negative answer follow-up panel can be added to increase text size of the html body using CSS. Further enhancements can be done in changing the structure of the website and navigation design of the website. It's preferable to have a separate UI framework to support dynamically changing information content of the website. In-order to distinguish each and every component of the website semantic HTML should be used in the User Interface Framework. HTML5 supports semantic elements and can be used for UI development.
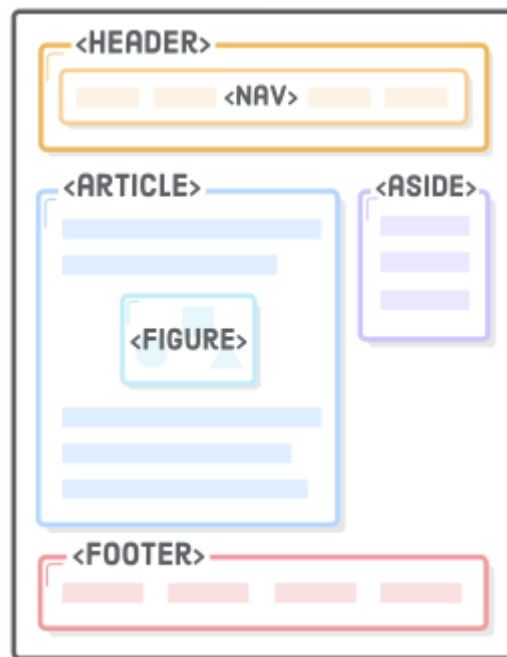

Figure 18: Semantic User Interface

If *<div>* elements were used for every component of the website it's an ambiguous structure and can't be identified properly. However using semantic UI as of Figure 18 each and every component can be identified. Imagine a situation where user prefers to view information content on the right side of the website. Normally most of the websites information contents are aligned to left since the heat map of most of the websites adhere to F- shaped pattern in reading, however according to Nielson it's just a flaw in design of the website and a good design can reduce F-pattern scanning of the website. So if the user prefers his/her information content to be on right rather than

being on left we can float the entire *<article>* to right by simply adding a style component to the *<article>* tag. To save these presets these changes can be saved as cookies or local storage of the user's web browser. Following figure 19 shows how semantic UI can be used in navigational design of a website.
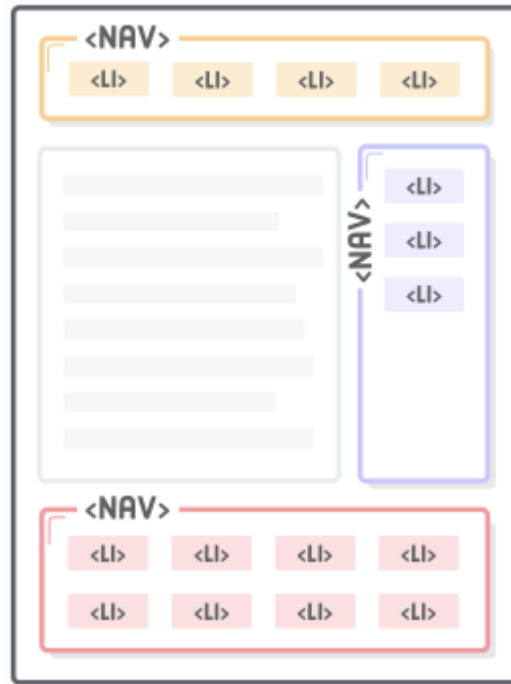
Figure 19 : Navigational semantic user interface

Imagine that user prefers a right navigational panel over navigational header or footer, user interface of the website can be easily changed using CSS style changes. So development of a semantic UI framework is preferable in incorporating personalization features to the website. However this personalization doesn't fall into the scope of this research. This research focus on usability evaluation through conversational interfaces which is used to identify what each users prefers.

Figure 20 : High-level diagram

CHAPTER 4

IMPLIMENTATION

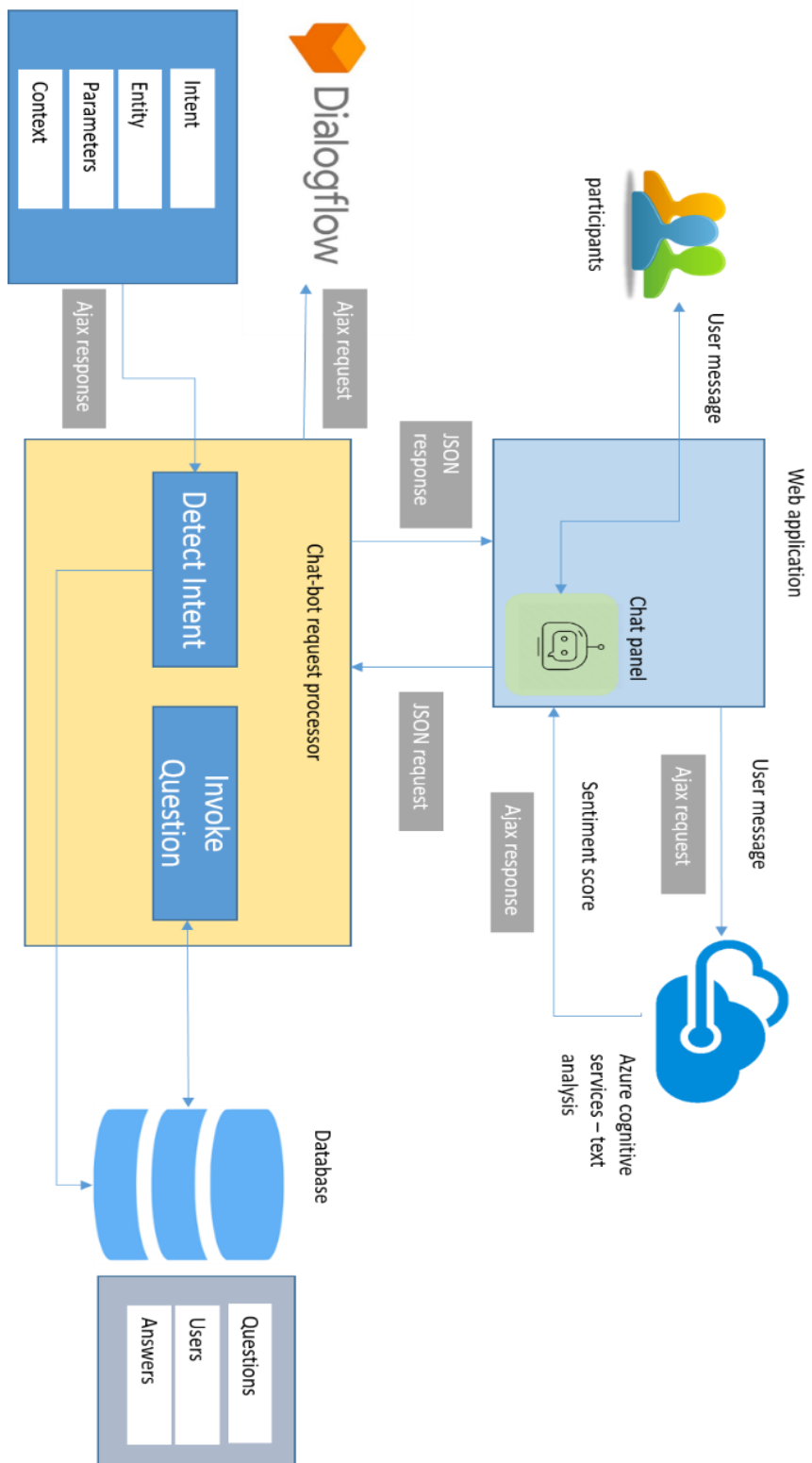## 4.1 Set the Goals of the usability evaluation

Before starting the evaluation exact goals of the evaluation should be identified, this will provide us the understanding of what you need to capture [36] during the evaluation. Setting goals before design gives the team a target to achieve. Following factors were based on their weightage identified on former studies of usability and user experience evaluation [37].

Selected factors,

- Usefulness
- Ease of Use
- Ease of Learning (Learnability)
- Satisfaction
- Attitude
- Attractiveness
- Navigational Structure
- Information Architecture
- Interface Design

These factors have been selected from Moville's user experience honey comb, Garrett's "Elements of User Experience", Laugwtz, Held and Schepp's evaluation and from the examination of questions from usability evaluation on questionnaires discussed on Chapter 3.

4.2 Determine the target population and sample size

The aim of the research is to provide a generalized set of usability and user experience evaluation questionnaire that could be integrated into any website as a Chatbot, so target population would be anyone using websites for every purpose. However in-order to carry-out the research a target population is selected and the procedure of selection of sample size is discussed on Chapter 5 along with a case study.

4.3 Determining the questions

We determined the goal of the research to capture, Usefulness, Ease of Use, Learnability, Satisfaction, Attitude, Attractiveness, Navigational structure, Information Architecture and Interface design. Questions were extracted from questionnaires and rephrased accordingly to form questions.

Table 4.3.i: Determining questions to be asked

Factors	Statements	Generated questions
Usefulness	It helps me be more effective. It helps me to be more productive	Does our website helps you to be more effective? Does our website helps you to be more productive? How our website useful in getting your work done?
Ease of Use	It is easy to use. It requires the fewest steps possible to accomplish what I want to do with it.	Do you find the website easy to use? tell me more about it Can you perform your tasks easily using the website?
Learnability	I easily remember how to use it. It is easy to learn how to use it.	Was it easy to learn?
Satisfaction	I'm satisfied with it. I feel the need to have it. It is fun to use.	Are you satisfied with the website? Do you feel need to have the website?

		Do you feel website is fun to use? How satisfied are you with the website?
Attitude	Using the system is a bad/good idea. The system makes work more interesting. I like working with the system.	How do you feel while using the system? Do you feel interested while using the system? What do you like or dislike when using the website?
Attractiveness	Attractiveness is captured through seven point scale from attractive to un attractive.	Do you find my interfaces boring or attractive?
Navigational Structure	-	What do you feel navigating with-in pages of the website?
Information Architecture	-	Do you think information content is correctly organized?
Interface Design	-	What color do you prefer as theme color?

4.4 Best practices in selection of better questions

4.4.1 Usage of simple language

Questions should be asked in a simple way without using any technical words or jargons, so it will be easy to be understood by the user. If the questions were ambiguous for the user he/she might not provide exact answer to the presented question which will lead to misinterpretation about the user experience of the website.

4.4.2 Avoid rhetorical questions

Rhetorical questions are used to giving implications of the answer or persuade the audience to comply. These questions are asked not hoping for an answer only to emphasize the correct answer. For an example questions like "The website's design is attractive, isn't it?" are rhetorical questions. The questions imply or persuade the audience to answer "yes, it's attractive". Since such questions won't provide users' honest answers and opinions about the website.

4.4.3 Include open-ended questions

Questionnaire should include close-ended questions to quantify the user feedback and evaluation of the website in certain dimensions such as Usefulness, Satisfaction, Ease of use and etc. However open-ended questions will reveal the actual issues of the user and it will give the evaluators an insight in to users' emotions while using the system.

4.4.4 Avoid invoking multiple questions

A question shouldn't be a combination of multiple questions. If a question contain multiple questions or statements it will be vague for the audience to answer since he/she will answer a question or provide multiple answers and we won't be able to maintain the context of the question. If more information is needed form a separate set of questions with required parameters to capture them.

4.4.5 Limit the questionnaire to few important questions.

Limiting the questions to be asked is tough decision to make since these questions will predict users' attitude towards the system is various dimensions. Most of the surveys with minimum number of questions tend to have higher completion rates and higher response rates [33]. However since we are using a Chatbot we could ask these questions over a period of time and map all the answered questions with the exact persona created.

4.4.6 Organize a flow in questions

The ultimate objective of the questionnaire is to capture more answers as possible. In order to do that the user should be kept in a longer conversation, for the questions should have a flow such as that we start with a set of simple questions and questions are asked in increasing order of complexity.

4.5 Conversational interface configuration

Dialogflow is used as the conversational interface to identify what users prefer to state based on their utterances. We have discussed about Utterances, Intent and Entities in Chapter 3. The basic conversation with the user and the Chatbot flows within following steps,

1. User typing the input.
2. Dialogflow agent parse the received input.
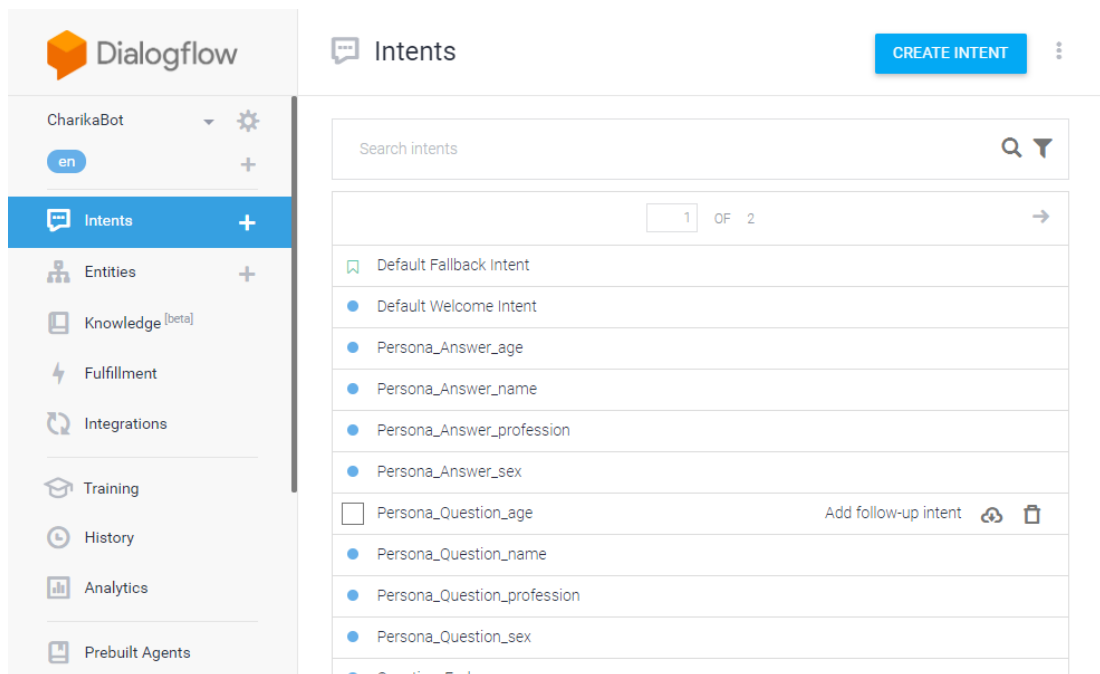3. Agent return the response to the user.

Figure 21: Dialogflow configuration

Here once we create an agent we can add intents to the created agent. Intents are responsible in mapping user input to its responses. In our model where we have Question and Answer. I have created Intent for each where question Intent will match to asked questions and Answer Intents will match to the answers provided by the users. However in general scenario Intent represents a single dialogue turn in a conversation. Intents consists of main components such as training phrases, Action and parameters

51

and responses. Training phrases are the utterance of the user or what people might say that matches our intent. These training phrases must be added in large scale in-order to train the Chatbot to identify various scenarios or instances of conversation. Parameters are information extracted from user utterances. For an Example if user will say "I'm 27 years old". By assigning a parameter age to the numerical value we can distinguish it from the utterance. A parameter can be a name, place, time or date and etc. Response is the utterance provided back to the user. You can configure what response is given back to the user. Let's look at how intents are configure in order to develop our Chatbot.
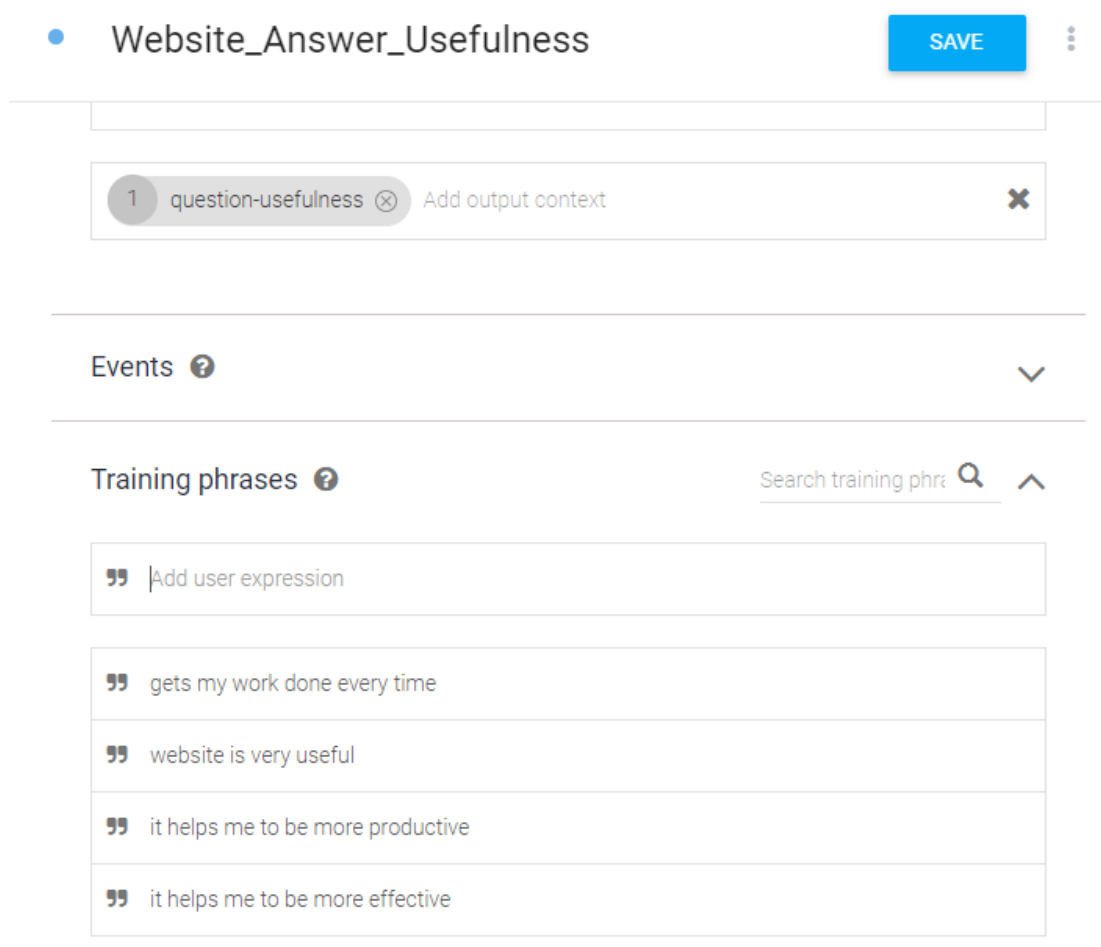

Figure 22 : Configure answer intent

The Figure 22 show the configured intent for the usefulness answer intent. Training phrases are added to detect intent based on user utterances. Contexts are used to match question and answers which we will discuss later.

Figure 23: Configure entities

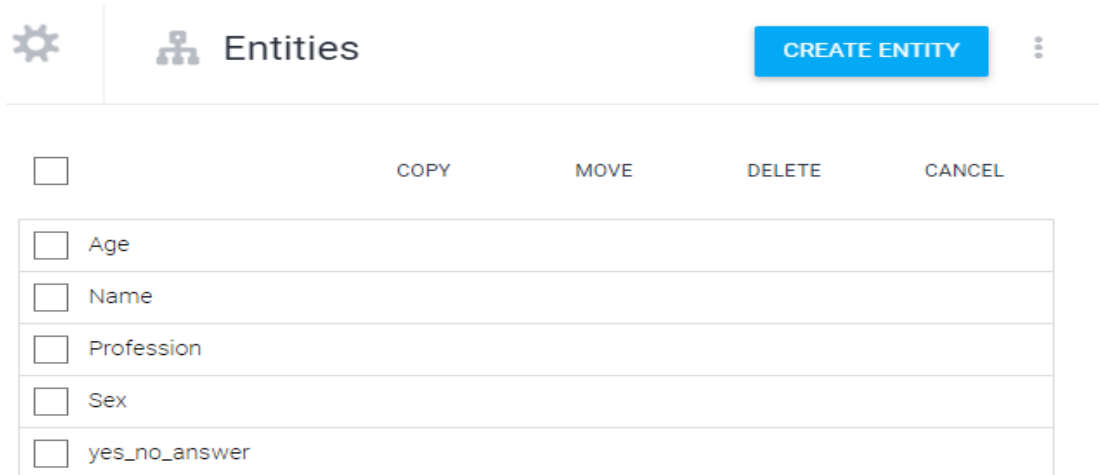These training phrases can be annotated by Entities. These annotations help Dialogflow to identify the exact intent. Entities represent real world entities. Which can be age, color, location and etc. parameters are used in identifying person questions to capture current user's age, profession and theme preference. Figure 24 shows how Persona_Answer_age intent training phrases are annotated with entities.

Figure 24: Persona answer configuration

Default Fallback Intent is used to respond to the user when answer provided by the user doesn't match with the Intents we have configured. In the conversational flow if the user's utterance won't match it will get the response from the fallback intent as "*I didn't get that. Can you say it again?*" Fallback intent is the backup mechanism in case of Chatbot failure, so the overall user experience won't be affected by Chatbot being not able to understand the user's input. In most of the Business application where Chatbot is unable to identify intention of the user, users will be redirected to a human agent maintaining the context of the conversation [38]. Having fallback intents are essential since classification of them reduces the risk of identifying it as another defined intent [39]. Imagine a situation where user asks about weather from the UX Chatbot. "How is the weather today in Colombo?" should not be mapped to a defined intent such as Persona_Answer_age Intent. So the fallback mechanism allows the classification of sentences that are not properly recognized by the Natural Language Understanding platform.

Now we have discussed how Intents are created, how training responses are added to train intent for all user utterances and how to annotate training responses based on entities. We mentioned how contexts involve in the conversation. Contents represent the current status of the users' intention and they help to carry information from one intent to another intent. This is how the conversation flows through. There are two types of contexts, Input contexts and Output contexts. Input contexts match when user utterance is a close match and if the context is active. Output contexts match only if the context is already matched. In our Question Answer scenario both Question and Answer is having the same output context so once questions is invoked output context is active and when the answer is provided by the user it will match with the output context of the Answer Intent. Multiple contexts can be assigned to an Intent however for this research purpose we are using only a single output context to have matching control in place. We have adjusted Lifespan of the output context to 1. Lifespan is the number of dialogue turns the context will remain active for. Since we are carrying out a usability questionnaire context of the question will remain until the user provides a suitable answer only. Once answer is captured context is removed.
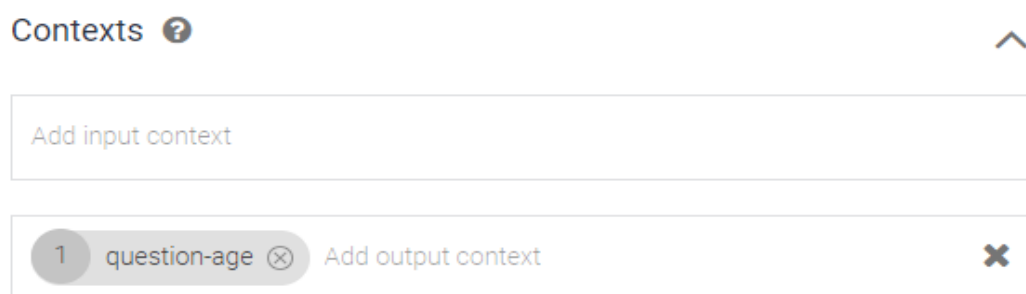

Figure 25: Configuring output contexts

4.6 Solution Architecture

We discussed about the High level diagram of the solution in Chapter 3. Figure 26 shows the solution architecture and its components. The application consists of a presentation layer where chat panel is integrated since we are using JavaScript and HTML the chat panel can be integrated any website.

Once chat panel is setup, the core functionality of processing users requests happen at Business logic processor. Chat panel checks whether user is typing or not in-order to invoke a new questions. Questions are also mapped in Dialogflow same as the answer. When invoke question method is executed it passes training phrases to retrieve the question. Questions are retrieved from unanswered questions by the user. Retrieved question is passed to the chat panel. Once a user inputs his/her message to the chat panel it is parsed to the BL processor with sentiment score captured through azure cognitive services text analysis service. BL processor will then detect the Intent by parsing it to dialogflow service. Dialogflow returns the identified Intent, Entities, Parameters and Contexts related to the passed user message. When valid answer is captured it will be saved in the database. These data will later be used in analysis dashboard. This is the core functionality of the system how questions get invoked and how the system captures the user answers.
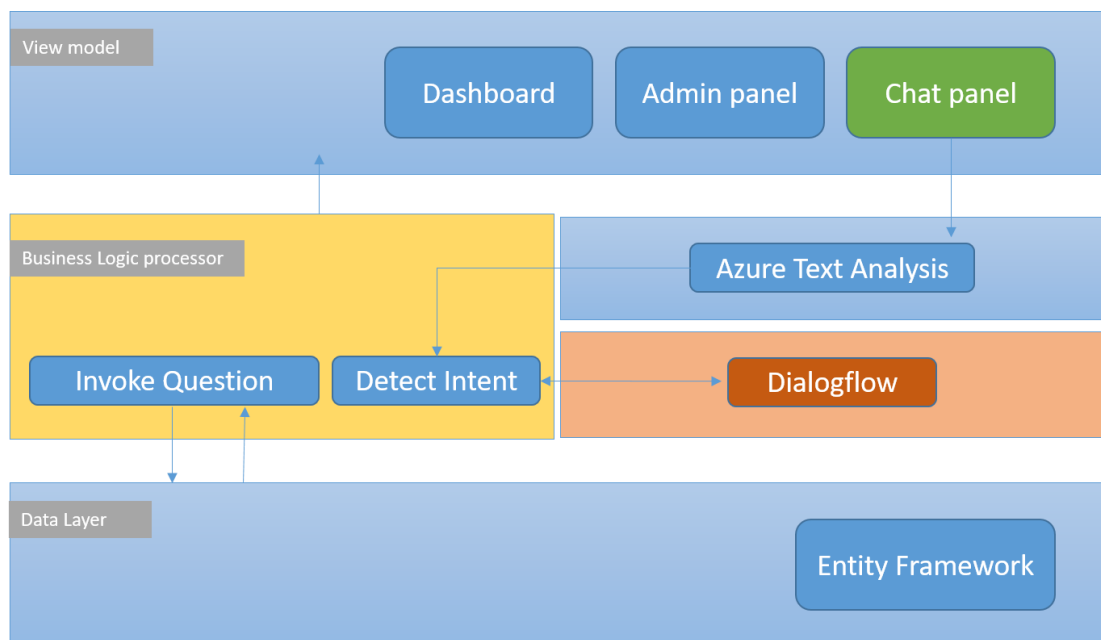

Figure 26 : Solution Architecture

4.6.1   Identifying new users and return users

In order to ask new questions from the new users and to continue asking remaining questions we need to maintain which user provides answers to the questions in user experience evaluation questionnaire. Also to identify personas from each other we need to maintain the current user. Since we have no access to register users before answering the questions we are gathering these information through the questionnaire and save these information in a cookie.

```javascript
// cookie values

function setCookie(name, value, days) {
    var expires = "";
    if (days) {
        var date = new Date();
        date.setTime(date.getTime() + (days * 24 * 60 * 60 * 1000));
        expires = "; expires=" + date.toUTCString();
    }
    document.cookie = name + "=" + (value || "") + expires + "; path=/";
    return getCookie(name);
}
function getCookie(name) {
    var nameEQ = name + "=";
    var ca = document.cookie.split(';');
    for (var i = 0; i < ca.length; i++) {
        var c = ca[i];
        while (c.charAt(0) == ' ') c = c.substring(1, c.length);
        if (c.indexOf(nameEQ) == 0) return c.substring(nameEQ.length,
c.length);
    }
    return null;
}

function eraseCookie(name) {
    document.cookie = name + '=; Max-Age=-99999999;';
}
```

Cookies are small piece of information saved in the users' web browser. Cookies are used by websites to remember stateful information. Persistent cookies are used with a specific date to expire. Cookie is set to expire on 365 days (1 year). This decision was made since it's identified that in order to have reliable information personas should be renewed within 1-4 years' time [23]. *setCookie()* method is used to set a cookie variable and expirer date. Once the given timeframe is exceeded cookie will be removed and this is identified by the Business Logic processor and it re-initiate persona questions to update the persona. *getCookie()* method is used to identify the

saved cookie value in users browser, such that we came to know that this is a returning user. *eraseCookie()* is used to erase a cookie stored in the users' browser.

A UUID is generated for each user and stored as a cookie. Users are distingue through this uuid. This is a hexadecimal number generated through Math.random function. Following method is executed once a new user initiates the chat.

```
function createUUID() {
    var s = [];
    var hexDigits = "0123456789abcdef";
    for (var i = 0; i < 36; i++) {
        s[i] = hexDigits.substr(Math.floor(Math.random() * 0x10), 1);
    }
    s[14] = "4";
    s[19] = hexDigits.substr((s[19] & 0x3) | 0x8, 1);
    s[8] = s[13] = s[18] = s[23] = "-";

    var uuid = s.join("");
    return uuid;
}
```

4.6.2 Invoking the question

Each question has been configured in Dialogflow with a training phrase, output context and a response. Also the Questions can be configured at Admin panel. When BL processor receive the invoke question executed it will retrieve questions that are not answered by the current user. Current user is identified by the cookie which we talked in the above section.

Figure 27: Question database table

```csharp
[HttpPost]
public JsonResult invokeQuestion(string user_id)
{
    string  questiontext = String.Empty;
    if (!bl_proc.IsUserAvailable(user_id))
    {
        bl_proc.AddUser(user_id);
    }
    //get unanswerd questions of the user
    Questions question = bl_proc.GetUnAnsweredQuestion(user_id);
    if (question.question_text != null)
    {
        questiontext = GetQuestion(question.question_text);
    }
    else
    {
        questiontext = GetQuestion("end the questions");
    }

    return Json(questiontext);
}
```

```csharp
public string GetQuestion(string questiontext)
{
    //set output context of the question
    Session["current_output_context"] = String.Empty;
    string projectId = "charikabot";
    string languageCode = "en-US";
    var client = SessionsClient.Create();
    var response = client.DetectIntent(
            session: new SessionName(projectId,
Guid.NewGuid().ToString()),
            queryInput: new QueryInput()
```

59

```
                   {
                       Text = new TextInput()
                       {
                           Text = questiontext,
                           LanguageCode = languageCode
                       }
                   }
               );

           //set output context of the question
           Session["current_output_context"] =
response.QueryResult.OutputContexts[0].ContextName.ContextId;

           dynamic queryResult = response.QueryResult;

           return queryResult.FulfillmentText;
       }
```

When invokeQuestion() method is called it will get user_id as an input and retrived unanswered questions of the current user from `GetUnAnsweredQuestion` method and question text is passed to dialogflow. Once the question invoking text is passed to dialogflow it will identify current question intent and return the response. Context of the question is also retrieved and stored in a session variable "`current_output_context`". This output context will be used to match a question intent with its exact answer intent.

4.6.3   Detecting the intent

Detection of the Intent is done through using Google Cloud Client libraries. Agent reference (*projectId*) and language of the user text (*LanguageCode*) is parsed as inputs with the user text in-order to identify the user intent.

```
string projectId = "charikabot";
string languageCode = "en-US";
var client = SessionsClient.Create();
var response = client.DetectIntent(
       session: new SessionName(projectId, Guid.NewGuid().ToString()),
                   queryInput: new QueryInput()
                   {
                       Text = new TextInput()
                       {
                           Text = text,
                           LanguageCode = languageCode
                       }
```

```
        }
    );
```

4.6.4 Analyzing the sentiment score

The questions in the questionnaire are open-ended questions and no leading questions were asked during the questionnaire session. So the answers users provide will be valuable insight to what users actually think about the system. In-order to capture rich responses from the user, user's sentiment is analyzed to understand whether their response is positive, negative or neutral.

To calculate the sentiment score Azure Text Analytics service is used. Azure Text Analytics is a cloud-based service which provides natural language processing from raw text. It's a part of azure cognitive services. This API is capable of processing the raw text provided and identify clues about positive or negative sentiment. It provides a sentiment score ranging from 0 to 1 for each document provided. Here 1 is the most positive while 0 being the most negative.

```
$.ajax({
        url:
"https://centralindia.api.cognitive.microsoft.com/text/analytics/v2.0/sentimen
t?" + $.param(params),
        beforeSend: function (xhrObj) {
            // Request headers
            xhrObj.setRequestHeader("Content-Type", "application/json");
            xhrObj.setRequestHeader("Ocp-Apim-Subscription-Key",
"a6143c127bde4a02a0413aa32699fa1a");
        },
        type: "POST",
        // Request body
        data: JSON.stringify(postData),
    })
    .done(function (data) {
        sentiment_score = data.documents[0].score;
});
    })
    .fail(function () {
        //
});
```

4.6.5   Matching output contexts of question and answer

Once the chat panel receives the exact question from Dialogflow it saves the current questions output context as a Session variable

```
//set output context of the question
Session["current_output_context"] =
response.QueryResult.OutputContexts[0].ContextName.ContextId;
```

Once the user provides an answer to the question in the chat panel intent detection function is called and check whether the current output context matches the output context of the answer. If it matches it is identified as an answer to the previously asked question.

```
//check answer match with available output contexts
foreach(var output_contexts in response.QueryResult.OutputContexts)
{
        if (Session["current_output_context"].ToString() ==
        output_contexts.ContextName.ContextId.ToString())
        {
                IsQuestionAnswered = true;
                Session["current_output_context"] = String.Empty;
                //if output contexts match each other save for former question
                bl_proc.AddUserAnswer(user_id,
output_contexts.ContextName.ContextId.ToString(), text, sentiment_score);

        }
}
```

Since an answer can be mapped with multiple intents we check running through all the contexts in a loop once both the output contexts are matched with each other it's identified as a valid answer to the question and it will be recorded to the database using Entity Framework.

	id	question_id	user_id	answer_provided	answer_intent	sentiment_value
37	54	10	ed53e20c-2fb5-43de-9c4e-c5d00e6a9cea	interface is very boring	NULL	0.005668491125106...
38	55	11	ed53e20c-2fb5-43de-9c4e-c5d00e6a9cea	I like the ordering process. and i dont like the...	NULL	0.05906844139099121
39	56	12	ed53e20c-2fb5-43de-9c4e-c5d00e6a9cea	yes it is properly organised	NULL	0.7888867855072021
40	57	13	ed53e20c-2fb5-43de-9c4e-c5d00e6a9cea	I prefer blue color	NULL	0.9445369243621826
41	58	1	e14e8619-4660-45cd-865c-49a412a44d...	yes	NULL	0.8335907459259033
42	59	2	e14e8619-4660-45cd-865c-49a412a44d...	i'm 27 years old	NULL	0.7519158720970154
43	60	3	e14e8619-4660-45cd-865c-49a412a44d...	i'm a software engineer	NULL	0.8511630296707153
44	61	4	e14e8619-4660-45cd-865c-49a412a44d...	navigating with-in website is ok	NULL	0.23193025588989258
45	62	5	e14e8619-4660-45cd-865c-49a412a44d...	male	NULL	0.5
46	1...	6	e14e8619-4660-45cd-865c-49a412a44d...	websie is useful in getting my work done	NULL	0.8641815781593323
47	1...	7	e14e8619-4660-45cd-865c-49a412a44d...	yes its very easy to use	NULL	0.9687134623527527
48	1...	8	e14e8619-4660-45cd-865c-49a412a44d...	since it's easy to rembermber i learnt it pretry ...	NULL	0.9861767292022705
49	1...	9	e14e8619-4660-45cd-865c-49a412a44d...	yes website is fun to use	NULL	0.9737439155578613
50	1...	10	e14e8619-4660-45cd-865c-49a412a44d...	your interfaces are very attractive and user fri...	NULL	0.9982566833496094
51	1...	11	e14e8619-4660-45cd-865c-49a412a44d...	i feel very disappointed while using the system	NULL	0.02488502860069275
52	1...	12	e14e8619-4660-45cd-865c-49a412a44d...	information content should be more oraganiz...	NULL	0.19320079684257507
53	1...	13	e14e8619-4660-45cd-865c-49a412a44d...	i like the black and white	NULL	0.8798261880874634
54	1...	1	61efc233-4567-4ce0-806b-a154ec14654c	yes	NULL	0.8335907459259033

Figure 28 : Question_Answer database table

4.6.6 Entity Framework database structure

Entity framework database first method is used to persist captured user personas, maintain questions to be asked and to capture answer for each question asked by the UI/UX questionnaire Chatbot. User Entity is used to persist persona, this will hold user's profession, age, sex, users will to participate in the questionnaire and his/her preferred theme in interface design. Question Entity holds the question text used to trigger the question from Dialogflow and the output context of the question (Question and answer is mapped using output context). Question_Answer will persist answers provided by the users per each question. It would contain the answer provided, sentiment value question_id of the questions and matched intent.
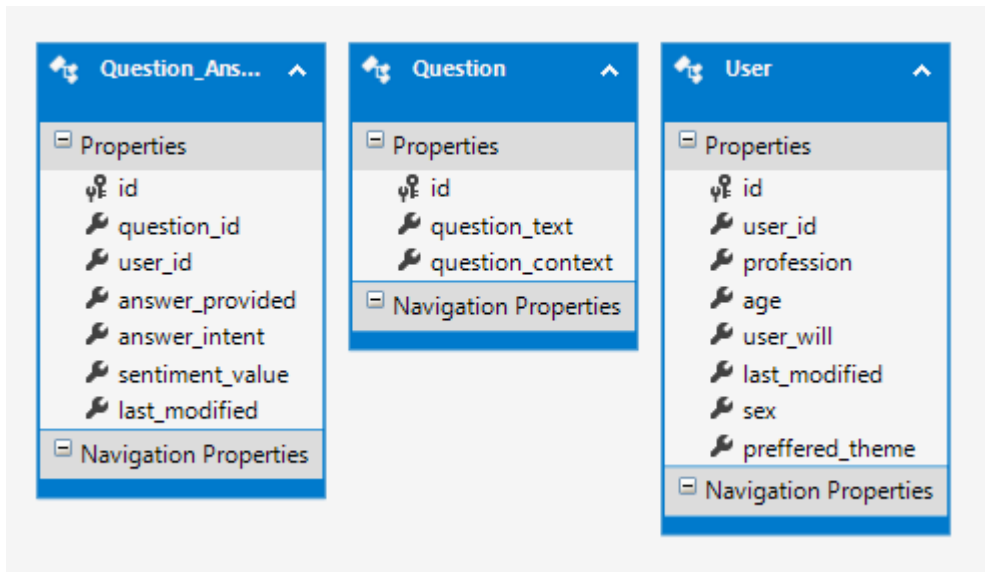
Figure 29 : Entity Framework Data model

4.7 Conversational user interface

The chat panel is placed at the bottom of the page where it doesn't affect the normal workflow of the system and where it is noticeable to the user. Figure 30 shows how chat panel is minimized and how it's less intrusive.
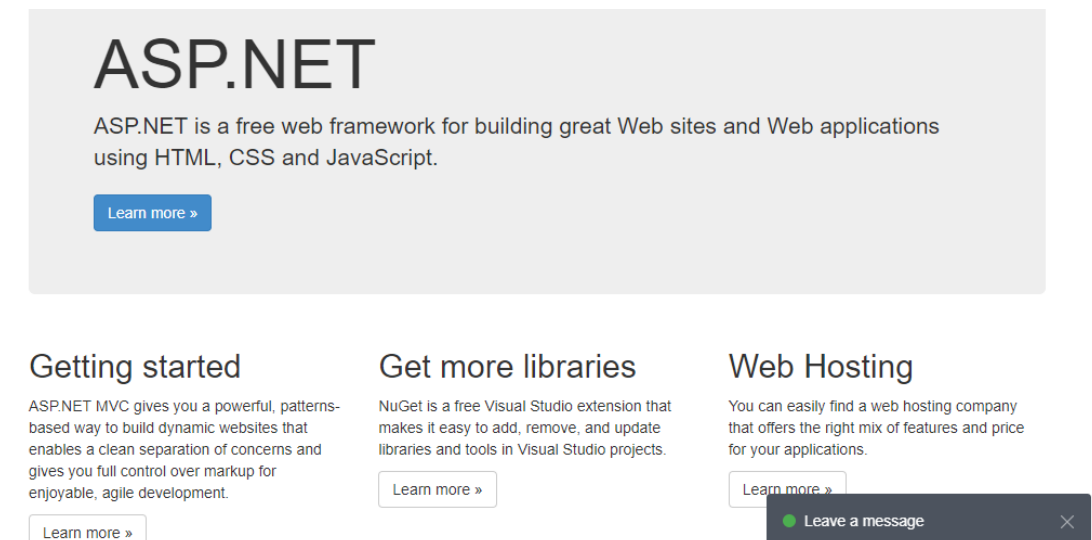

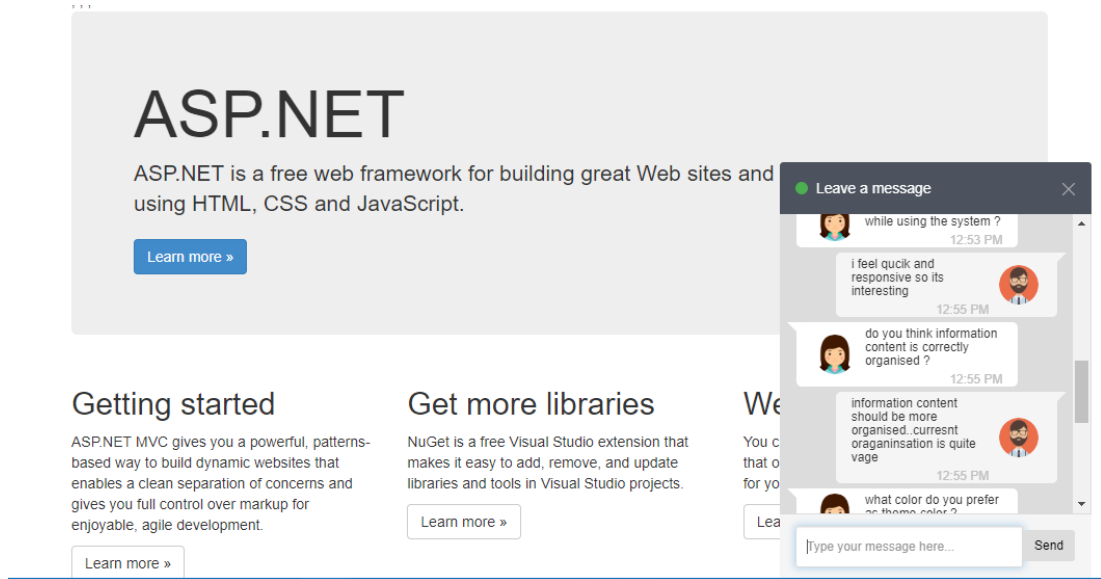*Figure 30: Chat panel minimized*

Figure 31 : Chatbot integrated to a test application

Chatbot interface is included with an avatar asking the questions such that the user feels that he/she is talking with a real human being in providing the answers to the questions.
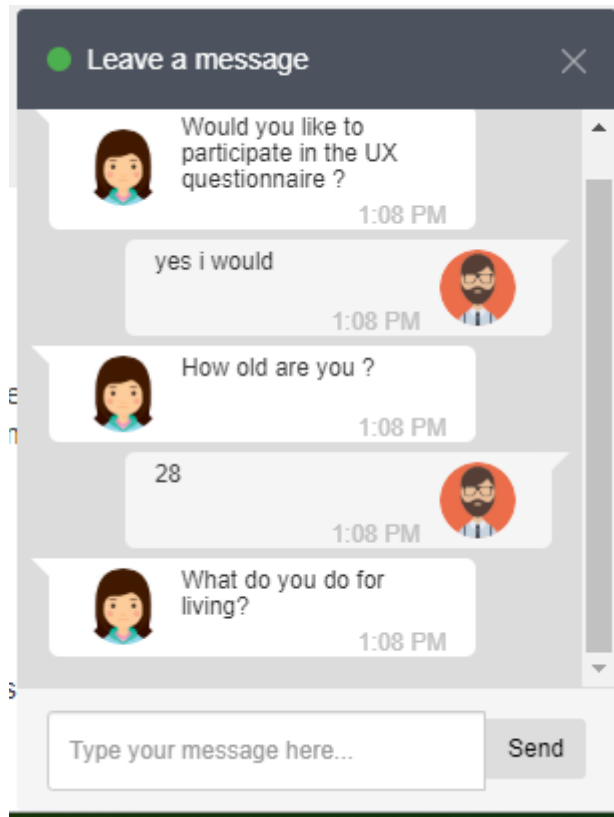

Figure 32: Chat panel with avatars

4.8 Pre-test questions

Questions were pre-tested using a Google Form. (Appendix A). These responses were used as training phrases in intent configurations.

4.9 Conduct questionnaire

Once pre-testing of the questions are done questionnaire is conducted to receive the exact user responses. Since the questionnaire is based on a Chatbot it is not distributed over target group of people via mail or printed media. Chatbot is integrated to the website using JavaScript and its HTML interfaces. Once it is added to the website Chatbot is ready go from that instance onwards.

4.9.1   Case Study – "CharikaBot"

"CHARIKA" is an initiative by SriLankan Airlines to handle staff tickets through a Booking engine in order to automatically issue tickets rather than manually capturing them through ticketing counters. "CHARIKA" the Self Service Portal for staff ticketing was introduced by SriLankan IT. CHARIKA Staff Travel System empowered employees, to easily reserve and efficiently manage staff and their eligible dependents leisure travel itineraries. Immediate payment through salary and online payment mode enabled instant ticket issuance any time anywhere. Productivity and efficiency was enhanced by direct integration with HR System which allowed to gather applicants', Payroll, Leave applications, retrieve employee dependent information with regard to employees. Easy reference of flight information and Real Time Booked Load, hassle free Self E-Ticket issuance through any device at any time

"CharikaBot" was integrated to the "CHARIKA" application and the user responses were gathered. A separate feedback questionnaire (Appendix B) was provided to the users in order to get users attitude towards the Chatbot questionnaire over traditional questionnaires.

4.9.2   Data representation

Once Chatbot is live, due to its non-invasive nature most of the "Charika" users were interested in trying it out for the first time. This interest supported the gathering of following data which provides insight on the usability of "Charika" website. The dashboard is used in admin panel to represent gathered data through the questionnaires. The Dashboard consists of Demographic information of the gathered data and Overall sentiment throughout the study. These sentiment values are average over gathered sentiment from 0 to 1 by sentiment analysis using Microsoft Text Analytics. Dashboard also contains a line chart which represents the variation of sentiment on each usability dimension over a period of time, this the implementation of emo-cards that we discussed over in Chapter 2 (Literature Survey), the main drawback of emo-cards were that they were based on the memory of the users that he/she draws the graph how the exact usability dimension changed over time in the specific website. In this implementation it will gather data based on the current view of the user and these variations are not based on memory.

Admin panel dashboard (Figure 33, 34) consists of an overall sentiment score graph. Which is sentiment of all user feedbacks throughout the questionnaire session on each usability factor. And it also consists of gender distribution pie chart based on responses from the users. However the authenticity of these data might not be accurate due to false information provided by the users due to distrust of the questionnaire. Overall sentiment graph with each usability factor sentiment variation over time shows how the users view on each usability factor changed over time. This is an exact representation of the UX curve [15, 16] which we discussed on chapter 2 (Literature Review) where users sketches on a curve how his/her relationship with the product changed over time based on his/her opinion.
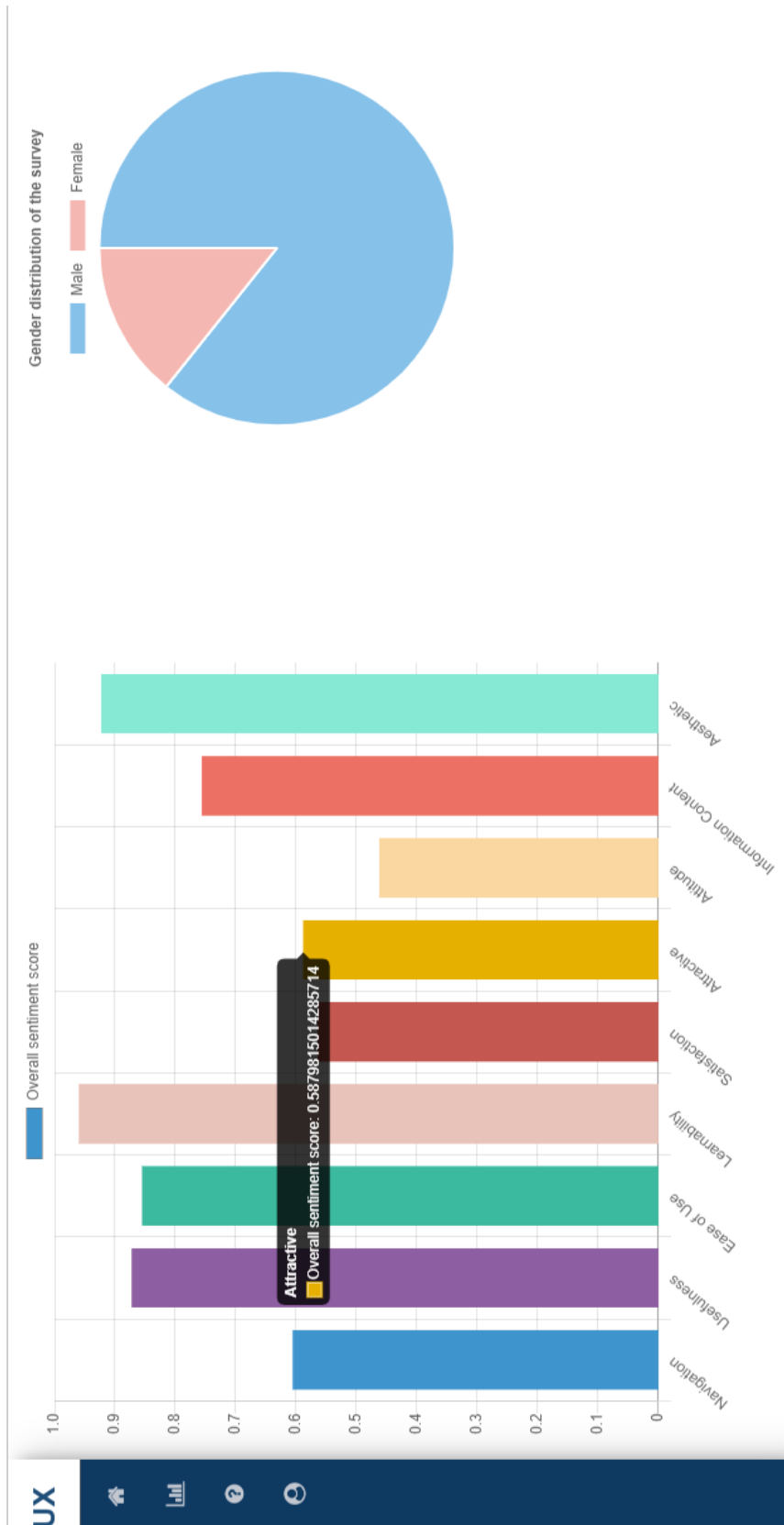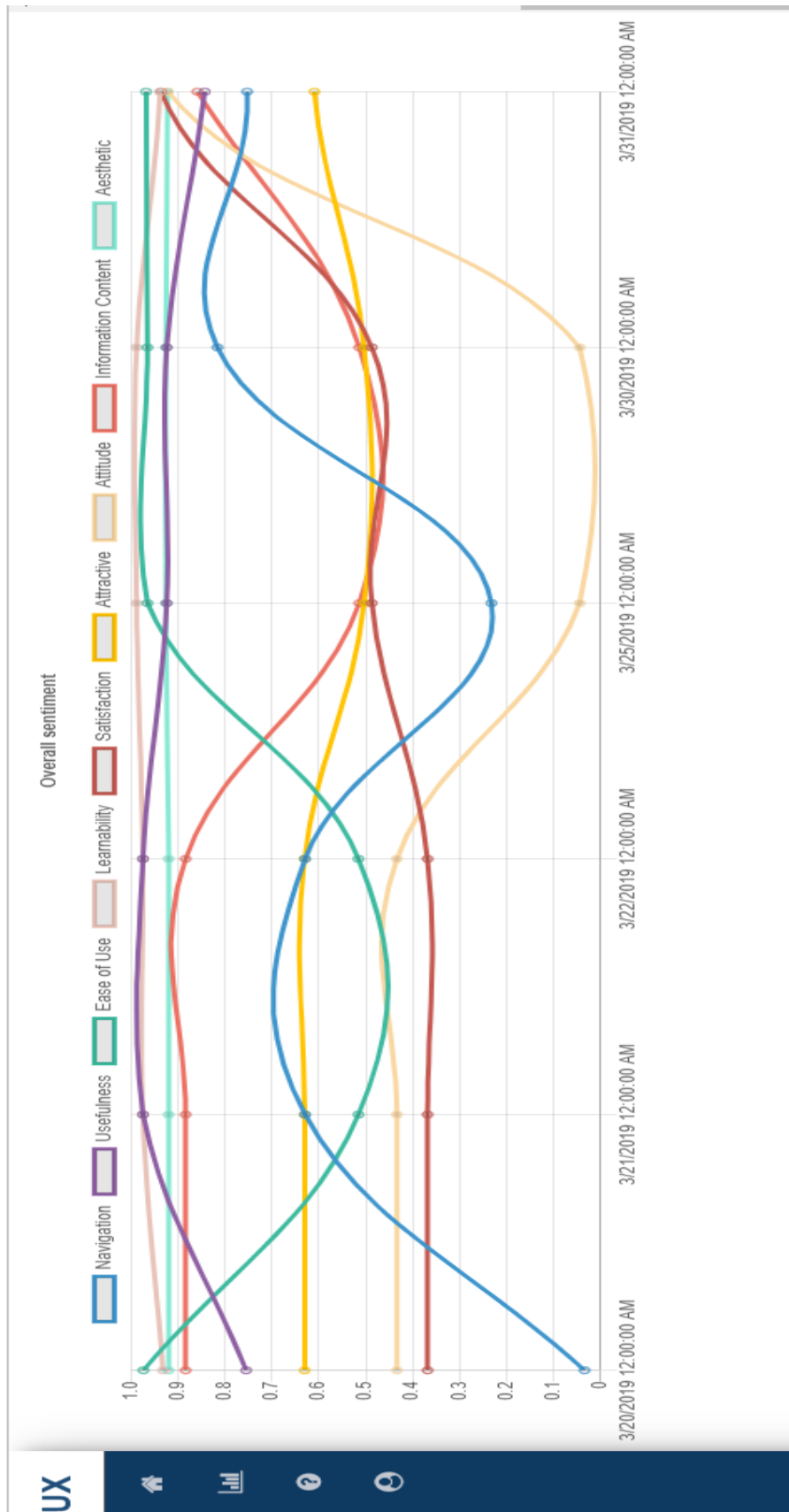
Figure 33: Overall sentiment score

Figure 34: Sentiment variation over time

Figure 35: Admin panel to update questionnaire

Figure 35 represents the Questions admin panel where we can change the Question asked, the question context which it maps with and usability dimension that each question is associated with. These configurations should be done with in synchronization with the Dialog flow training phrases and output contexts. Usability dimension is just to map each question with a usability factor where we can easily represent gathered data.

CHAPTER 5

EVALUATION

5.1 Performance Evaluation

Performance evaluation is carried out using a personal computer with following hardware configuration,

- Processor: Intel(R) Core(TM) i7-3630QM CPU @ 2.40GHz
- Random Access Memory: 4.00 GB
- Operating System: Windows 10 Enterprise, 64-bit Operating System
- Video Graphic Card: NVidia GeForce 610M.

Network connection:
- Ping: 32 ms
- Download speed: 9.34 Mbps
- Upload speed: 0.73 Mbps

Table 5.1.i : Task completion

Task	Time taken
Invoking a question	11 seconds 580 milliseconds
Retrieving sentiment score	1 second 15 milliseconds
Retrieving response	1 second 151 milliseconds

9.3% of responses from the users were not matched with any intent correctly. Once it is not matched with any intent Default Intent is fired in order to continue the conversation.
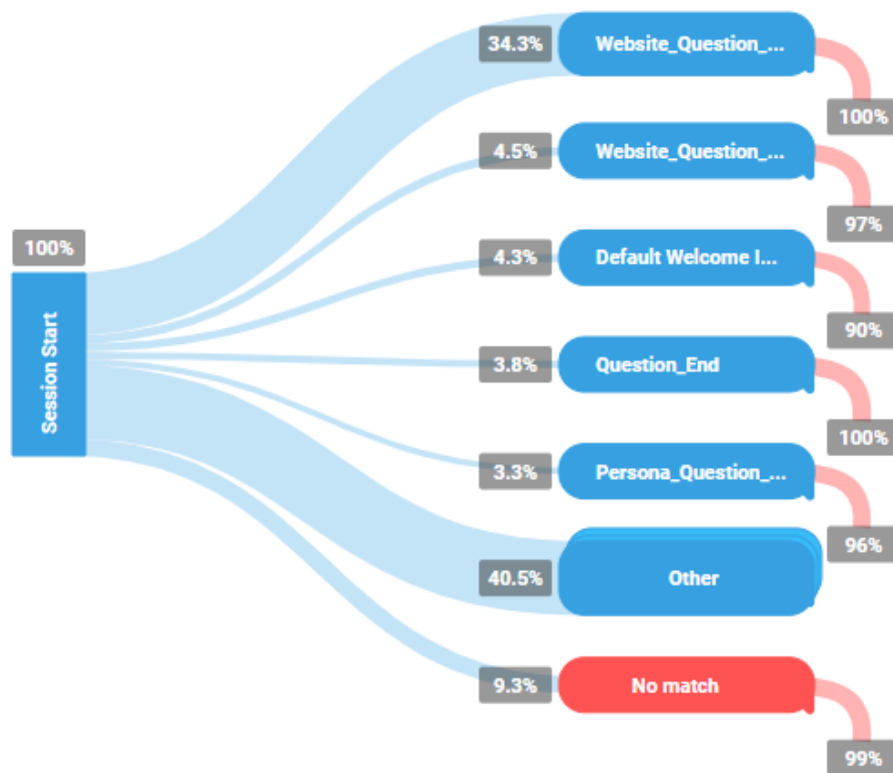
Figure 36: Conversational interface analytics

5.2 Heuristic Evaluation

Heuristic evaluation for the chatbot was carried out by experts including executives from UI/UX Engineer, Software Solution Architect, Business Analysts, Software Engineer and Quality Assurance Engineers. In order to evaluate the conversational interface Jakob Nielsen's Usability Heuristics were considered. Following were the concerns raised by the experts on "Charikabot".

Visibility of System status

System should always keep users about what is happening within the application

- It should provide feedback once user provides an answer to a question like "Good job...!"
- Chatbot should ask different questions when it can't understand what user said.

Match between system and real world

The system should use user's language which is familiar to the user. Don't use system oriented terms, use only natural conversation style.

- Some questions were leading and vague where user doesn't know what he/she should answer to the particular question.

User control and freedom

If user makes a mistake he/she can recover from errors and leave unwanted state without having to go through the process as is.

- User should be give control to skip questions

Consistency and standards

Users should not be left to wonder ambiguous or vague statements of the system. No concerns were raised since questions were quite understandable to every evaluator.

Error prevention

Without having user to recover from errors. Design the system carefully that system prevents the problem from occurring. In Chatbot we can prevent errors by having a good design in flow the conversation and having output contexts and input contexts to move from one dialogue occurrence to another. However it's very hard to prevent errors from appearing because most of the questions are expecting open-end answers where user can provide any query he/she think at the time. The evaluators came to the conclusion that proper fallback intent should be provided to recover where Chatbot can't understand the user's intent or to redirect them to a live chat where user can chat with a human being.

Flexibility and efficiency of use

Providing accelerators to the power users and start the conversations where the user left at.

No concerns were raised that Chatbot is capable of asking unanswered questions only. And it is capable of providing feedback that user has completed the survey without interrupting user every time he/she trying to access the system.

Aesthetic and minimalist design

Dialogues should not contain massive information content that are irrelevant or rarely needed.

- Proper animation when user and Chatbot is typing.
- Pop-up messages to get the attention of the user, given that it doesn't interrupt the user too much.

Help and documentation

Necessary help and documentation should be provided even system can be used without a documentation. Such information should be easy to found by the user

Experts raised the concern that "Charikabot" doesn't include any help or documentation. However some raised the concern that it would affect the humanly nature of the conversation.

- Should provide in-line help as in command line interfaces

5.3 User Feedback

Feedback was gathered from the users by conducting a questionnaire using a Google Form. The main objective of collecting user feedback was to understand how users think about using Chatbot instead of using online questionnaires for usability evaluation. This questionnaire (Appendix B) was provided to the users who have completed answering questions of the Chatbot.

Feedback questionnaire is based on Personality, Onboarding, Understanding, Answering, Navigation and Error Management.

Does the evaluator has clear understanding about the ongoing conversation ?

26 responses

● Yes
● No
● Maybe

7.7%

92.3%
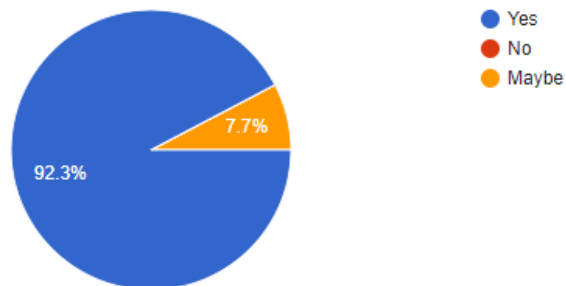
Figure 37: Feedback – Understanding Q1

92.3% rated that "Charikabot" is capable of understanding the context of the ongoing conversation. Chatbot Onboarding is to provide the users what its primary objective and goals. 80.8% of the users who provided feedback shows that onboarding task was successful.

Did he/she provide how to interact with him at the beginning ?

26 responses

● Yes
● No
● Maybe

11.5%

80.8%
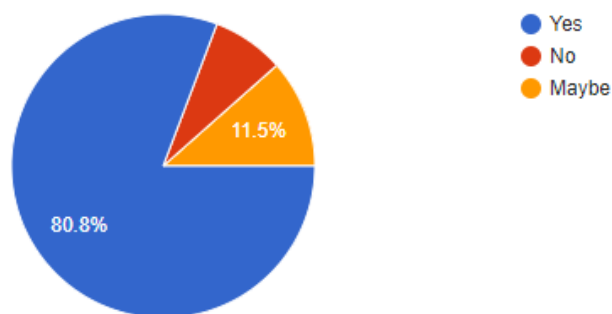
Figure 38: Feedback - Onboarding

Can he/she understand what you intend to do ?

26 responses
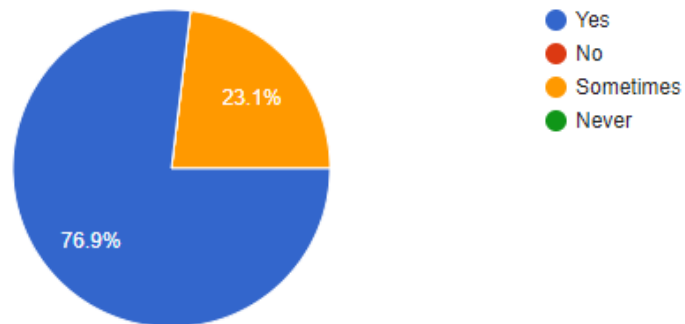
Figure 39: Feedback – Understanding Q2

76.9% rated that Chatbot is capable of understanding the user. While 23.1% rated it's capable of understanding the user sometimes. This is due to the fact that Chatbot is trained with Smalltalk and UI/UX questions only. In order to have a rich response mechanism much more intents should be added matching the common style of human conversation.

How much did you enjoy the providing answers to him/her ?

26 responses
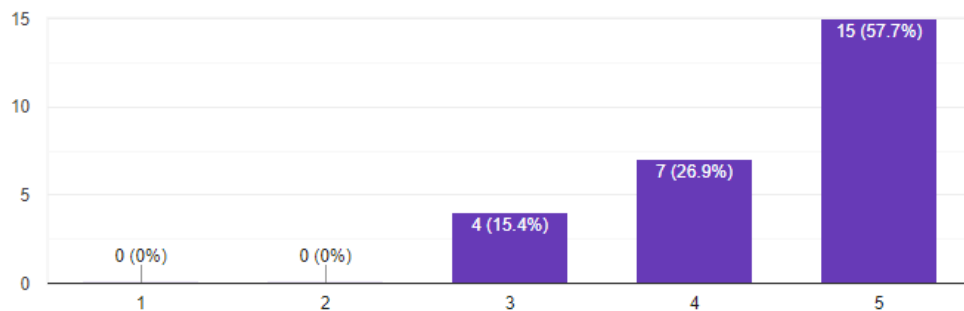
Figure 40: Feedback - Enjoyment rating

57.7% of the users strongly agreed, While 26.9% rated they agree while 15.4% had a neutral idea on that it's enjoyable to have a conversation with the Chatbot.

Do you feel lost while talking to the evaluator ?

26 responses
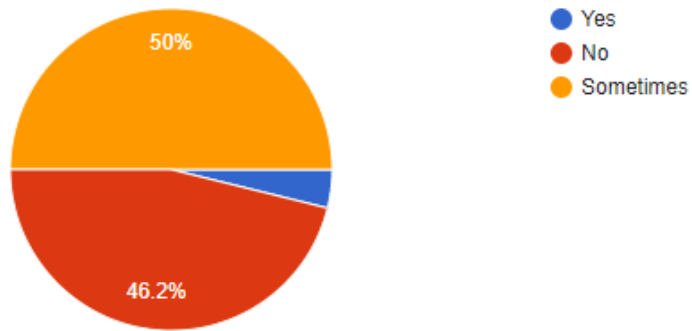
Figure 41 : Feedback - Navigation

50% rated they feel lost sometimes while Chatting with the Chatbot. Based on the discussion with the users, it's due to the fact that they need to provide an answer to the questions asked by the Chatbot. On Forms the answers are given and user just need to pick the correct answer which they think that is suitable.

Do you think he/she is capable of understanding you ?

26 responses
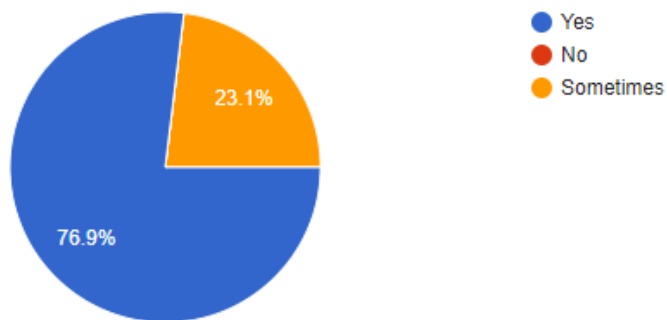
Figure 42: Feedback - Understanding Q3

10.0   10 (38.5%)

7.5   7 (26.9%)

5.0   5 (19.2%)

          4 (15.4%)
2.5

       0 (0%)
0.0
        1   2   3   4   5
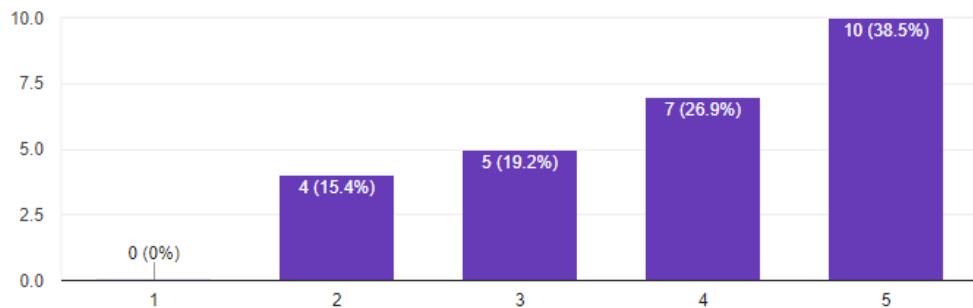
Figure 43 : Feedback - Chat interface rating

Chat interface is rated rating 5 by 38.5% of the users. And feedback showed chat interface need to be improved with emoji support. And concerns were raised to provide typing status animations to indicate that user or Chatbot is typing.

What would you rather choose between the chat and online form ?

26 responses

● Chat
● Form

26.9%

73.1%
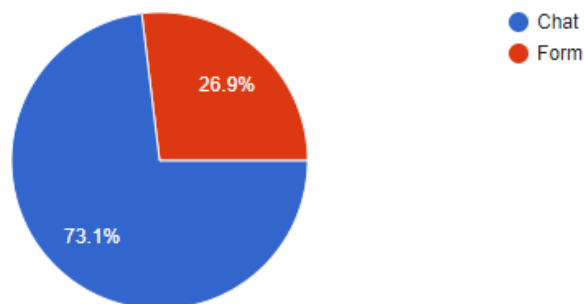
Figure 44: Feedback - User preference

73.1% rated that they choose Chat over Online Forms. This shows how Chatbots are effective in gathering data for usability evaluations. Only 26.9% agreed on the fact that forms are more desirable. Through the study we can state that Chatbot is more engaging that forms however it should support many intents as possible to mimic a real human being.

What did you think the evaluator is ?

25 responses

- A Human being
- Computer-based program
- I have no idea

20%

68%

12%
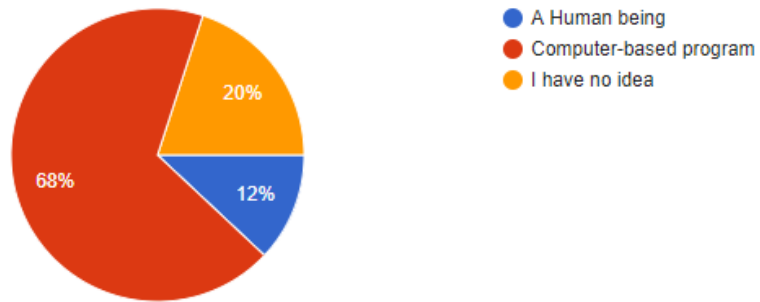
Figure 45: Feedback - Personality

The questionnaire was provided to the users after they completed answering the user experience questionnaire and they were not advised they are Chatting with a Chatbot. 68% of the users thought that the questions were asked from a computer based program, while 12% identified it as a real human being. Mainly the users involved to the case study was executives and their computer literacy is high which is reflected on the feedback as well they came to know that they are using a computer based program.

it was good chatting with someone rather than filling things up

conversations should be improved sometimes it cant understand what i say

providing answers was easy it didn't take much time to complete answering all the questions

sometimes it ask me the same question again. i don't want to answer it again

Chat interface can be improved adding emojis and animations on typing screen like in facebook

she should be more understand about user issue with the system. if she can provide answers about system issue its great

person chatting sometimes seems to be missing out my message

chatting is interesting than filling forms

Forms are easy to use than chatting.. chatting is time consuming

Chat should contain smileys like in facebook,whatsapp

chatting is okay..but ask too much personal information is bad

Forms are easy to fill. chatting takes time while at work

Figure 46: General feedback/Improvements

CHAPTER 6

CONCLUSION

This chapter provides a summary of the key points discussed in the research. It forms a narrative of the whole research thesis, incorporating ideas which have been constructed over the resea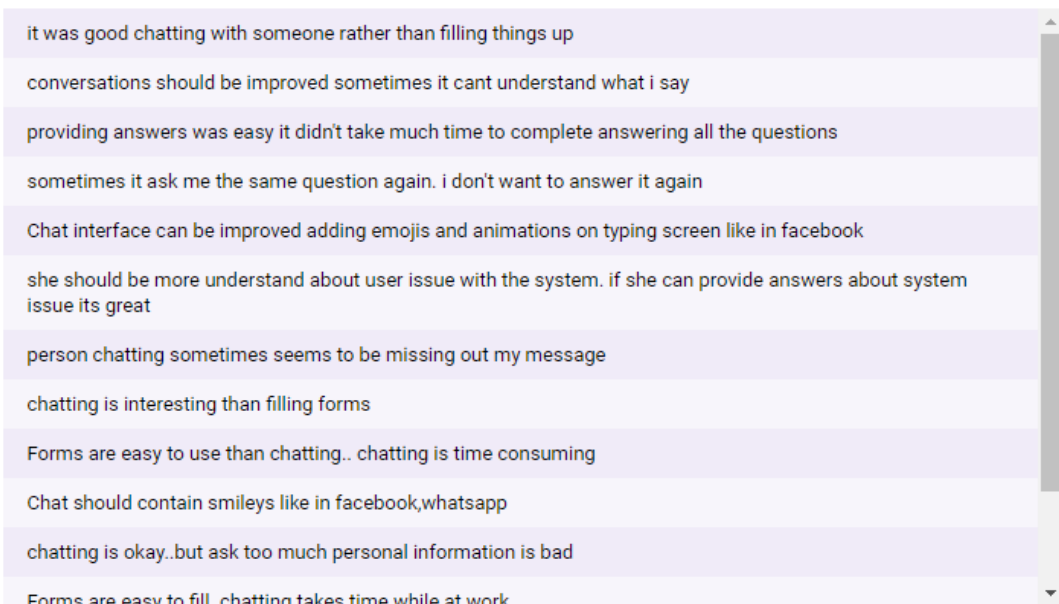rch study. At the end of this chapter where the research limitations and proposals for future research will be discussed

6.1 Summary of the Research

The aim of this research is to introduce a conversational interface for user experience evaluations which are currently done through questionnaires.

In Chapter 1 we formulated our research question which is "How can conversational interfaces improve quality of questionnaires for user experience evaluation, be more engaging and consistent in carrying out humanly conversations in-order to capture users attitude towards the system". Traditional questionnaires are set of questions with provided answers or Likert scores where users have to provide answers to, however, user's perspective on the system might be wholly different from answers that we provide. Also by providing statements for users to agree on we are forcing them to stick to an answer which is provided by questionnaires. Where we might not capture users exact attitude towards the system. The goal of the user experience design is to improve user satisfaction, loyalty and ease of use. To maintain constant user satisfaction throughout the product lifetime, conducting user experience evaluations are desirable however these evaluations tend to be short-term evaluations only focusing on initial product designs. The aim of the research is to use crowdsourcing mechanisms along with conversational interfaces to provide user experience evaluation throughout the products lifetime. We also focused on User experience and user experience evaluation, elements of user experience, crowdsourcing and persona creation.

In chapter 2, specific areas of previous research were examined referencing to the secondary statistical facts and reports of researchers investigated a topic related to this study. The literature review mainly focused on user experience evaluation techniques, crowdsourcing techniques, persona creation and conversational interfaces.

Chapter 3 was committed to present guidelines of research methodology followed throughout the research study, preferred technologies, technologies used and Software development methodology. It also provides the reasons behind the selection of preferable technology over the other. Chapter 4 provides the actual Implementation of the solution, How questions were determined, Configuration of conversational interface and Solution Architecture. Chapter 5 was intended to present data, gathered from the feedback survey from the users. In Chapter 6 the author will briefly conclude important findings in the research and discuss possible areas and questions for further research.

According to the chapter five data analysis, the final conclusion from the majority of the responses, Conversational interface (Chatbot) is preferable by most of the feedback providers, given that they are capable of clearly understanding the user. The humanly conversation was capable of gathering demographical data, improving the validity of the study where users tend to provide false information in traditional questionnaires. It is identified that the Chatbot should possess some intelligence over the daily activities of human beings in order to be more human as possible. (E.g. – ability to tell the current time) Even though fallback intents are preferable to overcome errors and keep up the conversation repetition of the fallback intent will discourage user from providing any answers.

6.2 Limitations

Identified limitations of the research are,

- Language support – the conversational interface currently supports only English. Even though it's implemented to support multiple languages research was carried out in English.
- Emoticons support – It only supports text queries, however in chatting people use emojis in a conversation and they are valuable in identifying user's emotion towards the system.

- User should have basic literacy in computing in order to participate in the conversation.

- Training phrases should be updated based on correct answers users provide, in order for the Chatbot to be intelligent and understand the majority of user intents.

- Providing multiple answers to a question with both negative and positive polarity will affect the sentiment score and provide false impressions on the system.

- A User providing simple answers like "yes", "no" won't be captured an answer for some questions in the questionnaire.

- Chatbot is not capable of identifying a feedback from the user without Chatbot initiating the question by itself.

- Limitation of available training preferences as all the usability evaluations are carried out through questionnaire likert scales.

This research study mainly focused on usability evaluation and usage of crowdsourcing to enhance the results of the usability evaluation. The preferences of the user are identified through the conversation with the Chatbot. These gathered information can later be used for improvement of the user interface. However through a UI framework and semantic user interface concepts a personalized interface can be provided to the user.

6.3 Future work

Future work will be increasing the Language support, converting all the intents to other languages where it supports many users from different geographic locations. This would increase the final output sentiment or overall user's attitude towards the system. Emoticons support should be included since most of the users will use emoji to express how they feel at the moment. Though emotion analysis we can support score provided though sentiment analysis. Providing intelligence to the Chatbot, new Intents should be added to support few useful functions like "What's the time?" and "What's the weather right now?" in order for the conversation to be humanly as possible. The

personality of the Chatbot should be increased where it can initiate a conversation on a topic which is not related to the questionnaire. Even though spammers are reduced through the context matching mechanism, more focus must be made on how to reduce "spammers" providing false information on demographic details. The scope of this research ends at gathering of sentiment scores on usability dimensions, understanding the user, his/her attitude future work should be done incorporating this gathered information to personalize content for each user of the website, where every user would have a personalized, optimized user experience with the website.

REFERENCES

[1] J. Garrett, *The Elements of User Experience*, 2nd ed. Berkeley (CA): New Riders, 2011, pp. 21-36.

[2] User Experience Design. [Blog] *Semantic Studios*.2004 Available at: http://semanticstudios.com/user_experience_design/ [Accessed 6 Nov. 2017].

[3] O. Gabry. (2016). UX—A quick glance about The 5 Elements of User Experience. [Blog] Available at: https://medium.com/omarelgabrys-blog/ux-a-quick-glance-about-the-5-elements-of-user-experience-part-2-a0da8798cd52 [Accessed 6 Nov. 2017].

[4] UX evaluation methods. [Blog] *ALL ABOUT UX*. 2015 Available at: http://www.allaboutux.org/all-methods [Accessed 6 Nov. 2017].

[5] T. Lavie. and N. Tractinsky. Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies*, 60(3), pp.269-298.2004

[6] K. Bang, M Kanstrup, A. Kjems, and J. Stage. *UX Evaluation Methods: An Investigation of the Danish IT-Industry's Work and the Relevance of Literature*. Master Thesis. Aalborg University - Department of Computer Science.2015.

[7] V. Roto , M. Obrist, and K. Mattila. User Experience Evaluation Methods in Academic and Industrial Contexts. In: *CHI2009 Conference*. Boston, Massachusetts: ACM Press.2009.

[8] G. Gediga, K. Hamborg and I. Düntsch, "The IsoMetrics usability inventory: An operationalization of ISO 9241-10 supporting summative and formative evaluation of

software systems", *Behaviour & Information Technology*, vol. 18, no. 3, pp. 151-164, 1999.

[9] M. Macleod, "Usability: Practical Methods for Testing and Improvement", in *Norwegian Computer Society Software 94 Conference*, Sandvika, Norway, 1994, pp. 1-12.

[10] "Software Usability Measurement Inventory", *Sumi.uxp.ie*. [Online]. Available: http://sumi.uxp.ie/en/. [Accessed: 11- Nov- 2017].

[11] B. Laugwitz, T. Held and M. Schrepp, "Construction and Evaluation of a User Experience Questionnaire", *Lecture Notes in Computer Science*, pp. 63-76, 2008.

[12] M. Rauschenberger, M. Schrepp, M. Perez-Cota, S. Olschner and J. Thomaschewski, "Efficient Measurement of the User Experience of Interactive Products. How to use the User Experience Questionnaire (UEQ).Example: Spanish Language Version", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 2, no. 1, p. 39, 2013.

[13] M. Manzoor,"Measuring user experience of usability tool, designed for higher educational websites". *Middle East Journal of Scientific Research*. 14. pp. 347-353.2013

[14] L. Hasan, "Evaluating the Usability of Educational Websites Based on Students' Preferences of Design Characteristics", *International Arab Journal of e-Technology*, vol. 3, no. 3, pp. 179-193, 2014.

[15] T. Koponen, J. Varsaluoma and T. Walsh, "Introduction to Long-Term User Experience Methods", Department of Software Systems Tampere University of Technology (TUT), 2011.

[16] S. Kujala, V. Roto, K. Väänänen-Vainio-Mattila, E. Karapanos and A. Sinnelä, "UX Curve: A method for evaluating long-term user experience", *Interacting with Computers*, vol. 23, no. 5, pp. 473-483, 2011.

[17] A. Kittur, E. Chi and B. Suh, "Crowdsourcing user studies with Mechanical Turk", *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, 2008.

[18] D. Liu, R. Bias, M. Lease and R. Kuipers, "Crowdsourcing for usability testing", *Proceedings of the American Society for Information Science and Technology*, vol. 49, no. 1, pp. 1-10, 2012.

[19]G. Meedin and I. Perera, "Crowdsourcing towards User Experience evaluation: An intelligent user experience questionnaire (IUEQ)", *2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2014.

[20] U. Kuter and C. Yilmaz, "Survey Methods: Questionnaires and Interviews", *Lte-projects.umd.edu*, 2001. [Online]. Available: http://lte-projects.umd.edu/hci-rm/survey.html. [Accessed: 06- Dec- 2017].

[21] W. Trochim, "Social Research Methods - Knowledge Base", *Socialresearchmethods.net*, 2000. [Online]. Available: https://www.socialresearchmethods.net/kb/. [Accessed: 06- Dec- 2017].

[22] "Ethical Guidelines for Statistical Practice", American Statistical Association, 2016.

[23] K. Flaherty, "Are Your Personas Outdated? Know When It's Right To Revise", *Nielsen Norman Group*, 2016.

[24] A. Affairs, "Personas | Usability.gov", *Usability.gov*. [Online]. Available: https://www.usability.gov/how-to-and-tools/methods/personas.html. [Accessed: 30- Nov- 2017].

[25] G. Ball, D. Ling, D. Kurlander, J. Miller, D. Pugh, T. Skelly, A. Stankosky, D. Thiel, M. Dantzich and T. Wax, "Lifelike Computer Characters: the Persona project at Microsoft Research", Microsoft Research, 2008.

[26] Nass, C.S., Jonathan; Tauber, Ellen R. Computers are Social Actors. In Proceedings of CHI'94: Human Factors in Computing Systems (April 24-28, Boston, MA), Association for Computing Machinery, 1994, pp. 72-77.

[27] B. Tidball and P. Stappers, "Crowdsourcing Contextual User Insights for UCD", 2011.

[28] "LUIS: Language Understanding Intelligent Service", *Luis.ai*. [Online]. Available: https://www.luis.ai. [Accessed: 06- Dec- 2017].

[29] J. Claridge, "WAMMI - Questionnaire", *Wammi.com*. [Online]. Available: http://www.wammi.com/questionnaire.html. [Accessed: 06- Dec- 2017].

[30] N. Eagle, "txteagle: Mobile Crowdsourcing", Lecture Notes in Computer Science, pp. 447-456, 2009.

[31] B. Ye, Y. Wang and L. Liu, "Crowd Trust: A Context-Aware Trust Model for Worker Selection in Crowdsourcing Environments", 2015 IEEE International Conference on Web Services, 2015.

[32] H. Yu, C. Miao, C. Leung, Y. Chen, S. Fauvel, V. Lesser and Q. Yang, "Mitigating Herding in Hierarchical Crowdsourcing Networks", Scientific Reports, vol. 6, no. 1, 2016.

[33] Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-analysis Sindre Rolstad, PhDlow asterisk,'Correspondence information about the author PhD Sindre RolstadEmail the author PhD Sindre Rolstad, John Adler, PhLic, Anna Rydén, PhD AstraZeneca R&D, Mölndal, Sweden

[34] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela and J. Weston, "Personalizing Dialogue Agents: I have a dog, do you have pets too?", Montreal Institute for Learning Algorithms, MILA, 2018.

[35] P. Mazaré, S. Humeau, M. Raison and A. Bordes, "Training Millions of Personalized Dialogue Agents", arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1809.01984. [Accessed: 2019].

[36] M. Dörflinger, A. Auinger and W. Wetzlinger, "Comparing Effectiveness, Efficiency, Ease of Use, Usability and User Experience When Using Tablets and Laptops", in International Conference of Design, User Experience, and Usability, University of Applied Sciences Upper Austria, Campus Steyr, 2014.

[37] "Usability Goal Setting Tool", IIT, Bombay.
[38] Everest Global, Inc, "Driving Enterprise Chatbot Adoption", 2018.

[39] M. Canonico and L. Russis, "A Comparison and Critique of Natural Language Understanding Tools", in The Ninth International Conference on Cloud Computing, GRIDs, and Virtualization, 2018, pp. 110-115.

[40] "Text Analytics API | Microsoft Azure", Azure.microsoft.com. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/. [Accessed: 01- Apr- 2019].

APPENDIX A

Usability evaluation of Charika

Please provide answers to following questions in order to provide you a better service

How old are you ?

Short answer text

Are you male or female ?

Short answer text

what is your designation ?

Short answer text

Usefulness

Description (optional)

Does our website helps you to be more effective ?

Short answer text

Does our website helps you to be more productive ?

Short answer text

How our website useful in getting your work done ?

Short answer text

Ease of use

Description (optional)

Do you find the website easy to use ? tell me more about it

Short answer text

Can you perform your tasks easily using the website ?

Short answer text

Ease of Learning

Description (optional)

Was it easy to learn ?

Short answer text

How did you remember how to use the website ?

Short answer text

Attractiveness

Description (optional)

Do you find my interfaces boring or attractive ?

Short answer text

what color do you prefer as theme color ?

Short answer text

Information Content

Description (optional)

What do you think about navigating with-in pages of the website ?

Short answer text

Do you think information content is correctly organised ?

Short answer text

Attitude

Description (optional)

How do you feel while using the system ?

Short answer text

Satisfaction

Description (optional)

How satisfied are you with the website ?

Short answer text

Do you feel website is fun to use ?

Short answer text

APPENDIX B

Evaluation of UI/UX evaluator

This questionnaire is focused on evaluation of conversation you had with UI/UX evaluator in Charika application. Please provide reliable answers to the questions. This would be a massive support for us to serve you better.

Does the evaluator has clear understanding about the ongoing conversation ?

○ Yes

○ No

○ Maybe

Did he/she provide how to interact with him at the beginning ?

○ Yes

○ No

○ Maybe

Can he/she understand what to intend to do ?

○ Yes

○ No

○ Sometimes

○ Never

How much did you enjoy the providing answers to his/her ?

	1	2	3	4	5
	○	○	○	○	○

Do you feel lost while talking to the evaluator ?

○ Yes

○ No

○ Sometimes

Do you thing he/she is capable of understanding you or tell you that he/she didn't understand ?

○ Yes

○ No

○ Sometimes

How would you rate the Chat interface ?

	1	2	3	4	5
	○	○	○	○	○

What would you rather choose between the chat and online form ?

○ Chat

○ Form

What did you think the evaluator is ?

○ A Human being

○ Computer-based program

○ I have no idea

Please leave feedback on your experience

ඔබේ පිළිතුර

submission 01

10	www.kurlander.net Internet Source	<%1
11	www.iaeng.org Internet Source	<%1
12	www.allaboutux.org Internet Source	<%1
13	www.nngroup.com Internet Source	<%1
14	academic.oup.com Internet Source	<%1
15	dl.lib.mrt.ac.lk Internet Source	<%1
16	www.dsoergel.com Internet Source	<%1
17	"Virtual, Augmented and Mixed Reality", Springer Nature, 2017 Publication	<%1
18	www.wseas.us Internet Source	<%1
19	open.uct.ac.za Internet Source	<%1
20	Ye, Bin, Yan Wang, and Ling Liu. "Crowd Trust: A Context-Aware Trust Model for Worker Selection in Crowdsourcing Environments",	<%1

2015 IEEE International Conference on Web Services, 2015.
Publication

21 "HCI and Usability for Education and Work", Springer Nature, 2008
    Publication   <%1

22 documents.mx   <%1
    Internet Source

23 www.iajet.org   <%1
    Internet Source

24 paperity.org   <%1
    Internet Source

25 David Squires, Jenny Preece. "Predicting quality in educational software:", Interacting with Computers, 1999
    Publication   <%1

26 Kujala, S.. "UX Curve: A method for evaluating long-term user experience", Interacting with Computers, 201109
    Publication   <%1

27 medium.com   <%1
    Internet Source

28 sycha.cz   <%1
    Internet Source

29 westus.dev.cognitive.microsoft.com
    Internet Source

$<\%1$

30 Celia Quico, Manuel José Damásio, Sara Henriques. "Digital TV adopters and non-adopters in the context of the analogue terrestrial TV switchover in Portugal", Proceedings of the 10th European conference on Interactive tv and video - EuroiTV '12, 2012
Publication

$<\%1$

31 minerva.leeds.ac.uk
Internet Source

$<\%1$

32 dspace.lboro.ac.uk
Internet Source

$<\%1$

33 El-Halees , Alaa | Abu-Zaid , Ibrahim M.. "Automated Usability Evaluation on University Websites Using Data Mining Methods = التقييم الآلي لسهولة استخدام مواقع الجامعات الإلكترونية باستخدام تقنية تنقيب البيانات", Palestinian Journal of Open Education, 2017
Publication

$<\%1$

34 arizona.pure.elsevier.com
Internet Source

$<\%1$

35 lasiweijiasiguojiyulechengguanfangwangzhi.aad246.top
Internet Source

$<\%1$